ISSN: 1859-2333

Tạp chí
KHOA HỌC
Trường Đại học Cần Thơ
Journal of Science, Can Tho University

Số chuyên đề
CÔNG NGHỆ THÔNG TIN
Special issue
Information Technology

Số chuyên đề (2013)
Special issue (2013)

Tạp chí Khoa học Trường Đại học Cần Thơ
Journal of Science, Can Tho University

ISSN: 1859-2333
Số chuyên đề (2013) - Special issue (2013)

Tổng Biên tập *(Editor-in-Chief)*
PGS. TS. Hà Thanh Toàn

Phó Tổng Biên tập *(Deputy Editor-in-Chief)*
PGS. TS. Nguyễn Thanh Phương

Thư ký Tòa soạn *(Editorial Secretariat)*
ThS. Trần Thanh Điện

Hội đồng biên tập *(Editorial board)*
GS. TS. Ngô Ngọc Hưng (Khoa học Đất)
GS. TS. Nguyễn Văn Thu (Chăn nuôi)
GS. TS. Lê Quang Trí (Quản lý Đất đai)
PGS. TS. Lý Nguyễn Bình (Công nghệ Thực phẩm)
PGS. TS. Nguyễn Duy Cần (Kinh tế - Xã hội)
PGS. TS. Võ Thành Danh (Kinh tế Nông nghiệp)
PGS. TS. Trần Nhân Dũng (Công nghệ Sinh học)
PGS. TS. Lê Việt Dũng (Khoa học Cây Trồng)
PGS. TS. Trần Cao Đệ (Công nghệ Thông tin)
PGS. TS. Trần Ngọc Hải (Nuôi trồng Thủy sản)
PGS. TS. Trần Thị Thanh Hiền (Nuôi trồng Thủy sản)
PGS. TS. Lê Văn Hòa (Khoa học Cây trồng)
PGS. TS. Bùi Thị Bửu Huê (Hóa học)
PGS. TS. Nguyễn Văn Nở (Ngôn ngữ học)
PGS. TS. Nguyễn Hiếu Trung (Quản lý Tài nguyên nước)
PGS. TS. Đỗ Văn Xê (Kinh tế Nông nghiệp)
GVC. TS. Thái Công Dân (Quản lý Giáo dục)
GVC. TS. Phan Trung Hiền (Luật Hành chính - Đất đai)
GVC. TS. Trần Văn Hiếu (Kinh tế Chính trị)
GVC. TS. Nguyễn Chí Ngôn (Tự động hóa)
GV. TS. Trần Trung Tính (Kỹ thuật điện)

Địa chỉ liên hệ
Tòa soạn Tạp chí Khoa học Trường Đại học Cần Thơ
Lầu 1, Khu Hiệu bộ, Khu 2, Đường 3/2, Quận Ninh Kiều, Thành phố Cần Thơ.
Điện thoại: *0710 3872 157;* ***Fax:*** *0710 3838 474*
Website: *sj.ctu.edu.vn;* ***Email:*** *tapchidhct@ctu.edu.vn.*

Chịu trách nhiệm nội dung
Tòa soạn Tạp chí Khoa học Trường Đại học Cần Thơ.
Giấy phép xuất bản số 200/GP.BTTTT của Bộ Thông tin và Truyền thông ngày 29/5/2013.

BAN BIÊN TẬP SỐ CHUYÊN ĐỀ:

CÔNG NGHỆ THÔNG TIN

Trưởng ban

PGS. TS. Trần Cao Đệ

Các thành viên

PGS. TS. Huỳnh Xuân Hiệp

TS. Trương Minh Nhật Quang

TS. Nguyễn Văn Hòa

TS. Ngô Bá Hùng

TS. Đỗ Thanh Nghị

TS. Phạm Nguyên Khang

TS. Lê Văn Lâm

TS. Nguyễn Thái Nghe

LỜI NÓI ĐẦU

Hơn 10 năm qua, Chính phủ đã thực hiện nhiều chủ trương, chính sách nhằm đẩy mạnh nghiên cứu và ứng dụng công nghệ thông tin (CNTT), trong đó xác định CNTT là một trong các động lực quan trọng nhất của sự phát triển, góp phần làm biến đổi sâu sắc đời sống kinh tế, văn hóa, xã hội của đất nước. Theo định hướng và chủ trương của Chính phủ, lĩnh vực CNTT đã phát triển mạnh ở các Trường và Viện nghiên cứu, trong đó Trường Đại học Cần Thơ trong những năm qua đã nhận được sự hỗ trợ tích cực của Bộ Giáo dục và Đào tạo, sự năng động hợp tác với các đơn vị trong và ngoài nước, đồng thời phối hợp với địa phương nghiên cứu và ứng dụng CNTT trong nhiều lĩnh vực.

Nhằm tiếp tục đưa CNTT vào đời sống xã hội, thúc đẩy phát triển kinh tế của quốc gia, Trường Đại học Cần Thơ tổ chức ***"Hội thảo Toàn quốc về Công nghệ Thông tin 2013"*** vào ngày 01 tháng 11 năm 2013 tại Trường. Hội thảo đã thu hút sự quan tâm của những nhà khoa học, nhà quản lý làm việc liên quan đến lĩnh vực CNTT trong và ngoài nước. Hội thảo có 75 bài viết dưới dạng tóm tắt, trong đó 62 bài chi tiết từ các nhà khoa học, giảng viên, nghiên cứu sinh và học viên đến từ các Viện/Trường của 13 tỉnh/thành đồng bằng sông Cửu Long, Thành phố Hồ Chí Minh, Bình Dương, Đà Lạt, Hà Nội, Hải Phòng và chuyên gia nước ngoài với 2 ngôn ngữ Việt và Anh. Tiểu ban Khoa học của Hội thảo đã chọn 23 bài viết xuất bản dưới dạng số chuyên đề (special issue) của Tạp chí Khoa học Trường Đại học Cần Thơ nhằm quảng bá thông tin đến độc giả.

Ban biên tập xin trân trọng giới thiệu đến quý độc giả Tạp chí Khoa học Trường Đại học Cần Thơ số chuyên đề về **Công nghệ Thông tin**.

TỔNG BIÊN TẬP

PGS.TS. Hà Thanh Toàn

MỤC LỤC

Xây dựng hệ thống thông tin chuyên ngành nông nghiệp phát triển nông thôn trên Webgis
Trần Lê, Trương Chí Quang, Lê Văn Thạnh, Võ Quang Minh và Phạm Văn Quỳnh *1*

Điều khiển robot Pioneer P3-DX bám sát đối tượng
Mã Trường Thành, Phạm Nguyên Khang, Lâm Chí Nguyện, Triệu Thanh Ngoan, Hà Duy An và Đỗ Thanh Nghị *11*

Giải thuật sửa lỗi cho các trình tự ADN ngắn
Nguyễn Văn Hòa và Nguyễn Văn Đông *20*

Thiết kế hệ thống thanh toán các dịch vụ công cộng ứng dụng công nghệ RFID và NFC
Đặng Vũ Minh Dũng, Đoàn Quốc Nam và Lương Vinh Quốc Danh *28*

Tổng quan về an ninh trên điện toán đám mây
Trần Cao Đệ *39*

Đánh giá kết quả học tập của sinh viên dựa trên tiếp cận trung bình trọng số kết hợp với trung bình trọng số được sắp thứ tự
Nguyễn Thị Thủy Chung và Huỳnh Xuân Hiệp *47*

Đa trừu tượng kết hợp tinh lọc trong kiểm tra mô hình hướng ký hiệu
Bùi Hoài Thắng và Nguyễn Thị Hồng Phấn *56*

Nhận dạng mã độc sử dụng cơ chế băm theo chỉ mục trên không gian dữ liệu phân hoạch
Trương Minh Nhật Quang *62*

Nâng cao độ chính xác phân loại lớp ít mẫu từ tập dữ liệu mất cân bằng
Bùi Minh Quân, Phạm Xuân Hiền và Huỳnh Xuân Hiệp *72*

CONTENT

Building agriculture and rural development database management system by WebGIS technology
Tran Le, Truong Chi Quang, Le Van Thanh, Vo Quang Minh and Pham Van Quynh *1*

Control of robot Pioneer P3-DX for object tracking
Ma Truong Thanh, Pham Nguyen Khang, Lam Chi Nguyen, Trieu Thanh Ngoan, Ha Duy An and Do Thanh Nghi *11*

Algorithm of short-read error correction
Nguyen Van Hoa and Nguyen Van Dong *20*

Design of an e-payment system for public services using RFID and NFC technologies
Dang Vu Minh Dung, Doan Quoc Nam and Luong Vinh Quoc Danh *28*

Overview on cloud computing security
Tran Cao De *39*

Evaluating student's study result based on weighted averaging combined with ordered weighted averaging approach
Nguyen Thi Thuy Chung and Huynh Xuan Hiep *47*

Multiple abstraction refinement in symbolic model checking
Bui Hoai Thang and Nguyen Thi Hong Phan *56*

Recognizing malicious code using hash indexing mechanism in categorical data space
Truong Minh Nhat Quang *62*

Improving prediction of the minority class in an imbalance dataset
Bui Minh Quan, Pham Xuan Hien and Huynh Xuan Hiep *72*

So sánh các mô hình dự báo lượng mưa cho Thành phố Cần Thơ
Đỗ Thanh Nghị, Phạm Nguyên Khang, Nguyễn Nhị Gia Vinh và Văn Phạm Đăng Trí 80

A comparision of rainfall forecast models for Can Tho city - Vietnam
Do Thanh Nghi, Pham Nguyen Khang, Nguyen Nhi Gia Vinh and Van Pham Dang Tri .. 80

Mô hình khảo sát ngưỡng đám đông trực tuyến và ứng dụng đề xuất một hệ thống đăng ký học phần mới cho Trường Đại học Cần Thơ trên nền tính toán đám mây với cấu hình hợp lý
Lê Quyết Thắng, Võ Hoàng Tú, Mai Yến Trinh và Võ Thị Cẩm Tú 91

Assessment survey modelling for the online crowd threshold problem applied to the course registration system in Can Tho University and recommending a new system based on the cloud computing with reasonable configuration
Le Quyet Thang, Vo Hoang Tu, Mai Yen Trinh and Vo Thi Cam Tu..... 91

Tiếp cận luồng cực đại trong mạng cho bài toán xếp lịch biểu
Phạm Nguyên Khang, Võ Trí Thức, Nguyễn Bá Diệp và Bùi Lê Diễm 100

Maximum flow network approach for scheduling problem
Pham Nguyen Khang, Vo Tri Thuc, Nguyen Ba Diep and Bui Le Diem..... 100

Khai khoáng và ứng dụng của mẫu episode mở rộng trên dữ liệu phụ thuộc thời gian
Nguyễn Bá Diệp, Phạm Nguyên Khang, Trần Nguyễn Minh Thư và Huỳnh Xuân Hiệp 109

Extended episode mining and its application in time-related data
Nguyen Ba Diep, Pham Nguyen Khang, Tran Nguyen Minh Thu and Huynh Xuan Hiep 109

Xây dựng mạng cảm biến không dây trong nông nghiệp chính xác
Lê Đình Tuấn và Thái Doãn Ngọc 115

Wireless sensor networks in agriculture
Le Dinh Tuan and Thai Doan Ngoc 115

Tư vấn học tập nhóm bậc đại học trên cơ sở tiếp cận lớp tương đồng lớn nhất trong hệ thống thông tin đa trị
Phan Tấn Tài, Trần Minh Tân, Phạm Ngọc Quyền, Nguyễn Thị Thu Thảo và Huỳnh Xuân Hiệp..... 123

Academic advising undergraduate groups based on the approach of maximal tolerance class of set-valued information system
Phan Tan Tai, Tran Minh Tan, Pham Ngoc Quyen, Nguyen Thi Thu Thao and Huynh Xuan Hiep 123

Giải thuật đơn giản để phát hiện làn đường và điều khiển lái cho ôtô tự hành
Trương Quốc Bảo 134

Simple lane detection and steering control algorithm for autonomous guided vehicle
Truong Quoc Bao 134

Theo dõi hiện trạng trà lúa phục vụ cảnh báo dịch hại lúa trên cơ sở sử dụng công nghệ viễn thám và hệ thống thông tin địa lý GIS
Trần Thị Hiền, Võ Quang Minh, Huỳnh Thị Thu Hương, Trần Thanh Dân, Hồ Văn Chiến, Nguyễn Hữu An và Nguyễn Phước Thành 143

Rice crop monitoring for early warning pest occurence using remote sensing and geographic information systems
Tran Thi Hien, Vo Quang Minh, Huynh Thi Thu Huong, Tran Thanh Dan, Ho Van Chien, Nguyen Huu An and Nguyen Phuoc Thanh 143

Một khung thức tìm kiếm và tái sử dụng hàm API tự động dựa trên đặc tả hình thức
Huỳnh Tấn Khải, Bùi Hoài Thắng, Quản Thành Thơ và Nguyễn Minh Thông..... *152*

A framework for automatic search and reusage of API functions based on formal specification
Huynh Tan Khai, Bui Hoai Thang, Quan Thanh Tho and Nguyen Minh Thong..... *152*

Xây dựng công cụ ngăn chặn truy việc cập web đen (hình ảnh, nội dung)
Huỳnh Bé Thơ và Trương Quốc Định..... *162*

Building tool to prevent access to black web (image, content)
Huynh Be Tho and Truong Quoc Dinh..... *162*

Xây dựng hệ thống gợi ý phim dựa trên mô hình nhân tố láng giềng
Triệu Vĩnh Viêm,Triệu Yến Yến và Nguyễn Thái Nghe *170*

Building a movie recommender system using factor in the neighbors approach
Trieu Vinh Viem,Trieu Yen Yen and Nguyen Thai Nghe..... *170*

Điều khiển tắc nghẽn trong giao thức truyền đa đường cho các ứng dụng multimedia
Lê Phong Dũ và Lê Tuấn Anh..... *180*

Congestion control in multipath protocol for multimedia applications
Le Phong Du and Le Tuan Anh *180*

Hệ thống gợi ý áp dụng cho trang web tổng hợp tin tức tự động
Đỗ Thành Nhân và Trần Nguyễn Minh Thư *190*

Recommender system for news aggregation website
Do Thanh Nhan and Tran Nguyen Minh Thu.. *190*

Mạng quan hệ giữa các đối tượng hình học và ứng dụng giải quyết bài toán vẽ hình tự động
Nguyễn Đình Hiển và Phạm Thi Vương..... *198*

Relational network of geometric objects and apply to draw the figure of problem automatically
Nguyen Dinh Hien and Pham Thi Vuong *198*

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

XÂY DỰNG HỆ THỐNG THÔNG TIN CHUYÊN NGÀNH NÔNG NGHIỆP PHÁT TRIỂN NÔNG THÔN TRÊN WEBGIS

Trần Lê[1], Trương Chí Quang[2], Lê Văn Thạnh[2],Võ Quang Minh[2] và Phạm Văn Quỳnh[1]

[1] *Sở Nông nghiệp và Phát triển Nông thôn Thành phố Cần Thơ*
[2] *Khoa Môi trường &Tài nguyên Thiên nhiên, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Building agriculture and rural development database management system by WebGIS technology

Từ khóa:
SharpMap, ASP.NET, SQL SERVER 2008, WebGIS, Google Map, Thủy lợi, Thủy sản, Thú y, Phát triển nông thôn, Bảo vệ thực vật, Cấp thoát nước

Keywords:
SharpMap, ASP.NET, SQL SERVER 2008, WebGIS, Google Map, Irrigation, Fisheries, Animal Health, Rural Development, Plant Protection, Water Supply

ABSTRACT

Building WebGIS system integrated database management provides the Department of Agriculture and Rural Development of Can Tho city data management (or call AGRIWEBGIS_CT) for the following six departments: Irrigation, Fisheries, Animal Health, Rural Development, Plant Protection, and Water Supply. AGRIWEBGIS_CTis built using ASP.NET Framework programming, database management system with new features in SQL SERVER 2008 version has an integrated spatial database and the open source library SharpMap. The system interfaces Vietnamese, many utility and ease of use as thematic maps to view. This system also provides aggregate statistics of each field data and option that allows system administrators to update the data of the departments. Besides, the general information page supports users during access as rural agriculture news.

TÓM TẮT

Xây dựng hệ thống WebGIS quản lý dữ liệu tổng hợp cho Sở Nông nghiệp và Phát triển Nông thôn Thành phố Cần Thơ (sản phẩm mang tên AGRIWEBGIS_CT) phục vụ cho quản lý dữ liệu của 6 Chi cục tương ứng với 6 phân hệ con (bao gồm: Thủy lợi, Thủy sản, Thú y, Phát triển nông thôn, Bảo vệ thực vật, Nước sạch và vệ sinh môi trường) và phân hệ quản lý tổng hợp. AGRIWEBGIS_CTđược xây dựng bằng phần mềm Microsoft Visual Studio.net có tích hợp ngôn ngữ lập trình VB.NET và CSharp, SQL SERVER với điểm mới ở phiên bản 2008 có tích hợp cơ sở dữ liệu không gian và thư viện GIS nguồn mở SharpMap. Hệ thống có giao diện tiếng Việt, nhiều tiện ích và dễ sử dụng như các bản đồ chuyên đề cho từng đơn vị. Đồng thời, hệ thống còn cung cấp các số liệu thống kê tổng hợp dữ liệu của từng lĩnh vực và cho phép cập nhật dữ liệu. Bên cạnh đó, các trang thông tin chung hỗ trợ người dùng trong quá trình truy cập như tin tức nông nghiệp nông thôn.

1 GIỚI THIỆU

Cần Thơ là thành phố trực thuộc Trung ương, nằm ở vị trí trung tâm của vùng đồng bằng sông Cửu Long và là vùng kinh tế trọng điểm thứ 4 của Việt Nam. Mặc dù mục tiêu chính của thành phố là đẩy mạnh công nghiệp hóa nhưng nông nghiệp vẫn đang giữ một vị trí rất quan trọng trong cơ cấu kinh tế - xã hội, thể hiện diện tích canh tác lúa ổn định trong khoảng 93.000-96.000 ha, sản lượng lúa ở mức 1,1-1,2 triệu tấn/năm; rau-màu khoảng 8.000

ha, sản lượng 96.000 tấn; cây công nghiệp ngắn ngày khoảng 5.000 ha, sản lượng 5.500 tấn; tổng đàn heo có khoảng 116.750, đàn gia cầm ước có 1.845.500 con, diện tích nuôi thủy sản đạt 12.550 ha, sản lượng 162.380 tấn (VIETRADE, 2012). Trong thời gian qua công tác thu thập, lưu trữ và quản lý dữ liệu nông nghiệp và phát triển nông thôn ở thành phố mặc dù được tổ chức, quản lý có hệ thống chặt chẽ nhưng vẫn còn gặp nhiều khó khăn do việc xử lý, lưu trữ theo hình thức giấy tờ truyền thống đã không tạo ra kết quả như ý muốn khi số liệu lưu trữ tích lũy theo thời gian và có sự thay đổi cán bộ thực hiện. Xuất phát từ những yêu cầu cấp thiết trên cùng với sự phát triển và hỗ trợ rất nhiều công cụ mạnh trong hệ thống thông tin địa lý GIS (Geographic Information System) kết hợp với kỹ thuật lập trình Web (còn gọi là công nghệ WebGIS) cho phép quản lý, chia sẻ, truy xuất và tìm kiếm dữ liệu hình học lẫn thuộc tính một cách dễ dàng (Ju & Chen, 2011; Zhang *et al*., 2008). Bên cạnh đó, với việc mở rộng của kiểu dữ liệu không gian (Spatial Data) trong SQL SERVER 2008 cho phép tích hợp kiểu dữ liệu Geometry để lưu trữ dữ liệu bản đồ chuyên ngành của các đơn vị thuộc Sở theo chuẩn dữ liệu không gian OGC (Open Geospatial Consortium). Theo Peter DeBetta *et al.* (2008) chuẩn không gian của kiểu dữ liệu này biểu diễn cho 7 dạng đối tượng như: Point (điểm), LineString (đường), Polygon (vùng), GeometryCollection (hỗn hợp), MultiPoint (nhiều điểm), MultiLine (nhiều đường) và MultiPolygon (nhiều vùng).

2 PHƯƠNG PHÁP NGHIÊN CỨU

2.1 Phương tiện thực hiện

Hệ thống được xây dựng trên nền Web ASP.NET, sử dụng ngôn ngữ Microsoft Visual Studio.NET, cơ sở dữ liệu được lựa chọn là SQL SERVER 2008 mà đơn vị đã được trang bị từ các dự án trước đây, thư viện GIS nguồn mở SharpMap được lựa chọn do hỗ trợ cơ sở dữ liệu không gian lưu trên SQL SERVER đồng thời SharpMap là công nghệ thuận lợi trong quản lý kết hợp dữ liệu không gian và thuộc tính.

Nguồn dữ liệu không gian và thuộc tính của các chi cục và trung tâm: Bản đồ hành chính; dữ liệu và bản đồ chuyên ngành từng Chi cục.

2.2 Phương pháp nghiên cứu

2.2.1 Phương pháp thu thập dữ liệu

Thu nhập cơ sở dữ liệu bản đồ nền hành chính thành phố từ Sở Tài nguyên và Môi trường. Thu thập các lớp bản đồ và số liệu chuyên đề cho 6 phân hệ quản lý từ các Chi cục, trung tâm thuộc Sở Nông nghiệp và Phát triển Nông thôn như: bảo vệ thực vật, thú y, nước sạch và vệ sinh môi trường, phát triển nông thôn, thủy lợi, thủy sản.

2.2.2 Chuẩn hóa dữ liệu không gian

Nguồn dữ liệu bản đồ thu thập được gồm nhiều định dạng (MapInfo, MicroStation, AutoCAD, tọa độ GPS). Sử dụng phần mềm MapInfo tiến hành chuyển đổi định dạng, nắn chỉnh và chuyển các lớp bản đồ không cùng hệ quy chiếu về một hệ quy chiếu với lưới chiếu WGS84, vùng lưới chiếu là UTM Zone 48-Northern Hemisphere.

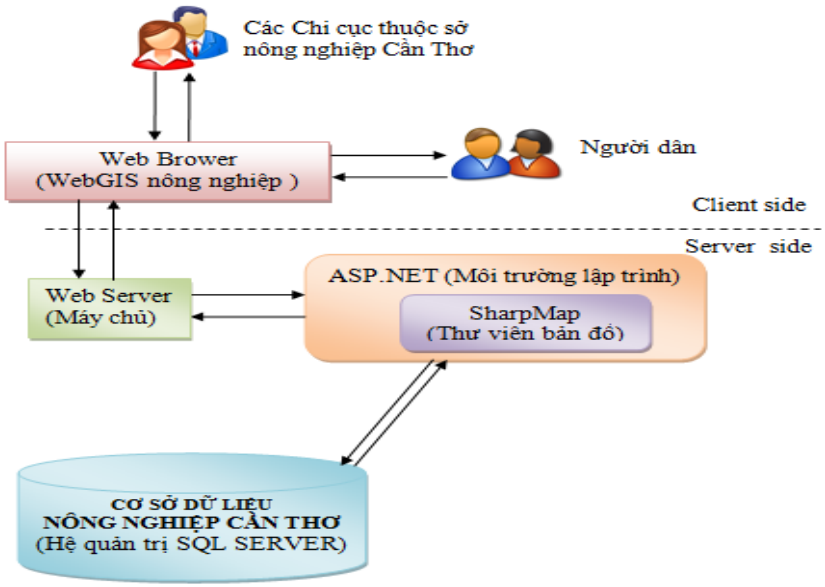
2.2.3 Phương pháp cơ sở dữ liệu

- Chuyển dữ liệu không gian đã biên tập từ dạng Shapefile vào trong CSDL SQL SERVER 2008.
- Xây dựng CSDL thuộc tính và không gian trên SQL SERVER 2008.
- CSDL của hệ thống cho 6 phân hệ được tách ra thành 7 CSDL tập trung trên 1 máy chủ bao gồm cơ sở dữ liệu nền dùng chung và 6 CSDL tích hợp không gian và thuộc tính của 6 phân hệ.

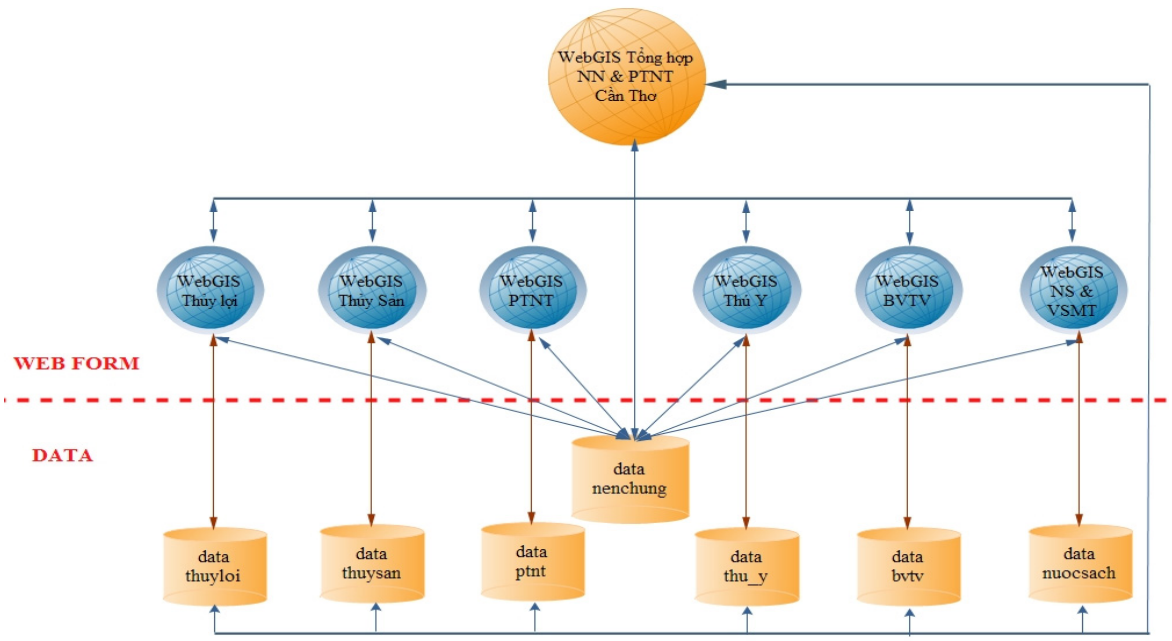
2.2.4 Xây dựng các ứng dụng GIS tích hợp trên Website

Sử dụng phương pháp thân tích hệ thống cho từng phân hệ quản lý dữ liệu của 6 đơn vị triển khai ứng dụng GIS của Sở bao gồm: Thủy Lợi, Thủy Sản, Phát triển nông thôn, Thú Y, Bảo vệ thực vật, Nước sạch và Vệ sinh môi trường.

Kiến trúc hệ thống được triển khai như sơ đồ Hình 1, trong đó sử dụng thư viện bản đồ nguồn mở SharpMap như một máy chủ dịch vụ bản đồ (Web Map Service - WMS) nhằm tận dụng khả năng truy vấn và quản lý dữ liệu hiệu quả của bộ thư viện này.



Hình 1: Kiến trúc hoạt động của hệ thống



Hình 2: Mô hình kết nối giữa các các phân hệ với cơ sở dữ liệu của hệ thốngAGRIWEBGIS_CT

3 KẾT QUẢ VÀ THẢO LUẬN

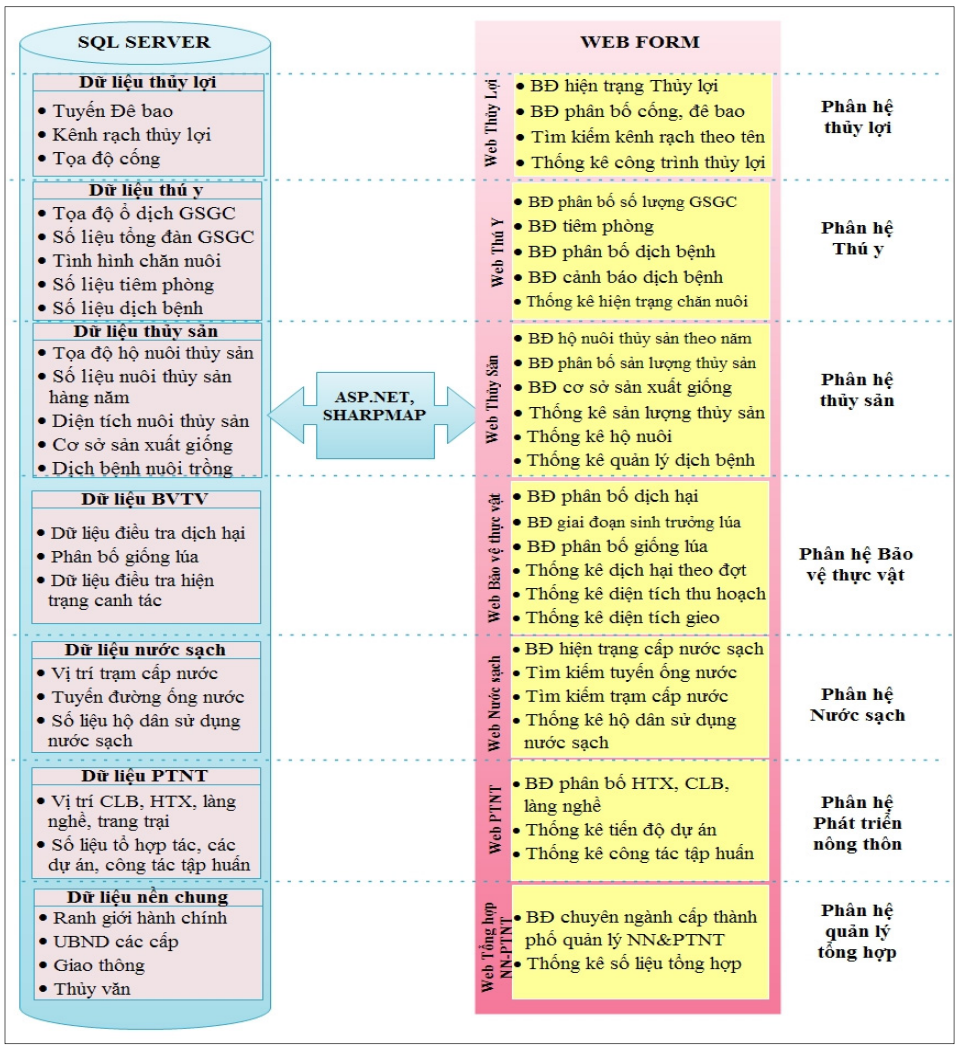
3.1 Mô hình liên kết giữa các phân hệ quản lý trong hệ thống AGRIWEBGIS_CT

Mô hình ứng dụng của AGRIWEBGIS_CT được xây dựng như Hình 2 với một hệ thống gồm 7 khối chức năng tương ứng với 7 phân hệ quản lý. Trong đó, phân hệ thứ nhất quản lý dữ liệu tổng hợp nông nghiệp và phát triển nông thôn của thành phố (hay còn gọi phân hệ cha), còn 6 phân hệ còn lại tương ứng với 6 Chi cục quản lý trực thuộc Sở (hay gọi phân hệ con), các phân hệ này truy xuất đến CSDL nền dùng chung và CSDL dành riêng cho phân hệ. Cơ sở dữ liệu riêng bên trong mỗi phân hệ con chứa các số liệu chuyên ngành, các bản đồ chuyên ngành và cùng liên kết với dữ liệu nền chung hành chính của thành phố (data nenchung). Mỗi phân hệ con này được lập trình phát triển tạo thành một hệ thống Web Form để hiển thị thông tin và truy vấn dữ liệu chuyên ngành

của mỗi Chi cục đang quản lý. Phân hệ cha cho phép liên kết đi đến giao diện trang chủ Web Form của 6 phân hệ con đồng thời cũng được thiết kế các chức năng truy vấn nhanh các bản đồ và các báo cáo của các phân hệ con nhằm giúp cán bộ quản lý ở Sở quản lý được thông tin cập nhật tức thời.

3.2 Quản lý dữ liệu chuyên ngành của mỗi phân hệ

Hệ thống AGRIWEBGIS_CT cho phép đa người dùng truy xuất đồng thời trên nền web. Mỗi phân hệ cung cấp các thông tin chuyên ngành của lĩnh vực mà Chi cục đang quản lý dưới dạng thống kê số liệu và bản đồ chuyên đề trực tuyến. Đồng thời, chức năng cập nhật dữ liệu cũng được xây dựng kèm theo cho mỗi phân hệ giúp cho người quản trị hệ thống dễ dàng thêm mới, quản lý và chỉnh sửa dữ liệu. Hình 3 minh họa các chức năng chính và dữ liệu lưu trữ trong các CSDL của các phân hệ.



Hình 3: Mô hình mô tả thông tin chuyên ngành của các phân hệ trong AGRIWEBGIS_CT

Phân tích chức năng từng phân hệ cho việc xây dựng ứng dụng cho các Chi cục quản lý thuộc sở như sau:

3.2.1 Phân hệ quản lý thủy lợi:

– Nhập dữ liệu đê bao: Cho phép vẽ mới đê bao và nhập dữ liệu thuộc tính của đê.

– Nhập dữ liệu trạm bơm: Cho phép vẽ thêm trạm bơm bằng tọa độ x, y theo phép chiếu UTM và nhập dữ liệu thuộc tính của trạm bơm.

– Kết xuất các báo cáo thống kê số lượng và chiều dài kênh thuỷ lợi, đê bao theo các cấp tỉnh huyện xã theo năm hay tại thời điểm thống kê.

– Xây dựng bản đồ trực tuyến: Bản đồ hiện trạng các đê bao, kênh rạch và công trình cống.

3.2.2 Phân hệ quản lý dữ liệu Thú y

– Nhập liệu: Cho phép nhập dữ liệu ổ dịch: tạo điểm ổ dịch theo bằng tọa độ x, y theo phép chiếu UTM.

– Thống kê dữ liệu: Báo cáo dịch bệnh theo cấp hành chính theo tháng, năm.

– Bản đồ trực tuyến: Xây dựng bản đồ tổng hợp số liệu tổng đàn loại gia súc và gia cầm; bản đồ vị trí cơ sở kinh doanh thức ăn, thuốc thú y, thu gom chế biến trứng.

– Thông tin bệnh: Thông tin về các loại dịch bệnh ở gia súc và gia cầm. Nguyên nhân gây ra dịch bệnh, triệu chứng của bệnh, biện pháp phòng và trị bệnh.

3.2.3 Phân hệ quản lý dữ liệu phát triển nông thôn

– Nhập liệu: xây dựng bản đồ trang trại, HTX, làng nghề ở dạng điểm bằng tọa độ x, y theo phép chiếu UTM và nhập các số liệu.

– Xây dựng chức năng tìm kiếm – thống kê: Báo cáo công tác, tổng hợp dự án, thông tin HTX, thông tin làng nghề.

– Xây dựng chức năng hiển thị bản đồ: Bản đồ phân bố các HTX, làng nghề, trang trại.

3.2.4 Phân hệ quản lý nước sạch

– Chức năng nhập liệu: Nhập dữ liệu tổng hợp về cấp nước cho hộ dân nông thôn, dữ liệu không gian và thuộc tính cho các trạm cấp nước, dữ liệu đường ống chính nối với các trạm cấp nước.

– Xây dựng chức năng tìm kiếm – Báo cáo: Thông tin trạm cấp nước, trạm cấp nước, đường ống cấp nước.

– Xây dựng chức năng hiển thị bản đồ: Trạm cấp nước theo cấp quản lý, bản đồ chuyên đề số lượng hộ dân nông thôn sử dụng nước sạch theo cấp hành chính.

3.2.5 Phân hệ quản lý bảo vệ thực vật

– Xây dựng chức năng nhập liệu:Nhập dữ liệu điều dịch hại, dữ liệu canh tác và giống lúa của địa phương.

– Xây dựng chức năng tìm kiếm – Báo cáo: Báo cáo tình hình dịch hại theo đợt, diện tích gieo trồng và thu hoạch.

– Xây dựng chức năng hiển thị bản đồ: Bản đồ phân bố giống lúa, bản đồ phân bố dịch hại, bản đồ giai đoạn sinh trưởng.

3.2.6 Phân hệ quản lý dữ liệu thủy sản

– Xây dựng chức năng nhập liệu: Nhập hộ nuôi thủy sản, dữ liệu nuôi trồng hàng năm, nhập danh mục loài thủy sản.

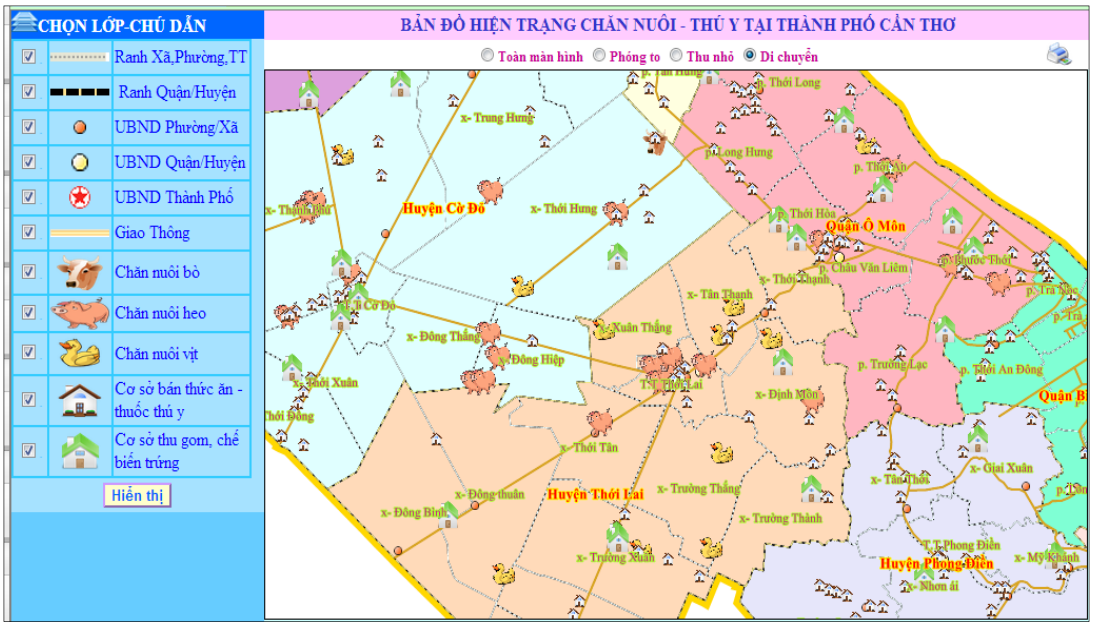
– Xây dựng chức năng tìm kiếm – Báo cáo: sản lượng, diện tích, số hộ nuôi theo tháng, năm, báo cáo sản lượng của từng kích cỡ từng loại thủy sản cung cấp thông tin giúp cho doanh nghiệp thu mua, đơn vị quản lý.

– Xây dựng chức năng hiển thị bản đồ: Bản đồ các hộ nuôi thủy sản theo năm, bản đồ sản lượng thủy sản/ năm/đơn vị hành chính.

3.3 Chức năng của AGRIWEBGIS_CT

3.3.1 Chức năng thao tác bản đồ trực tuyến các phân hệ

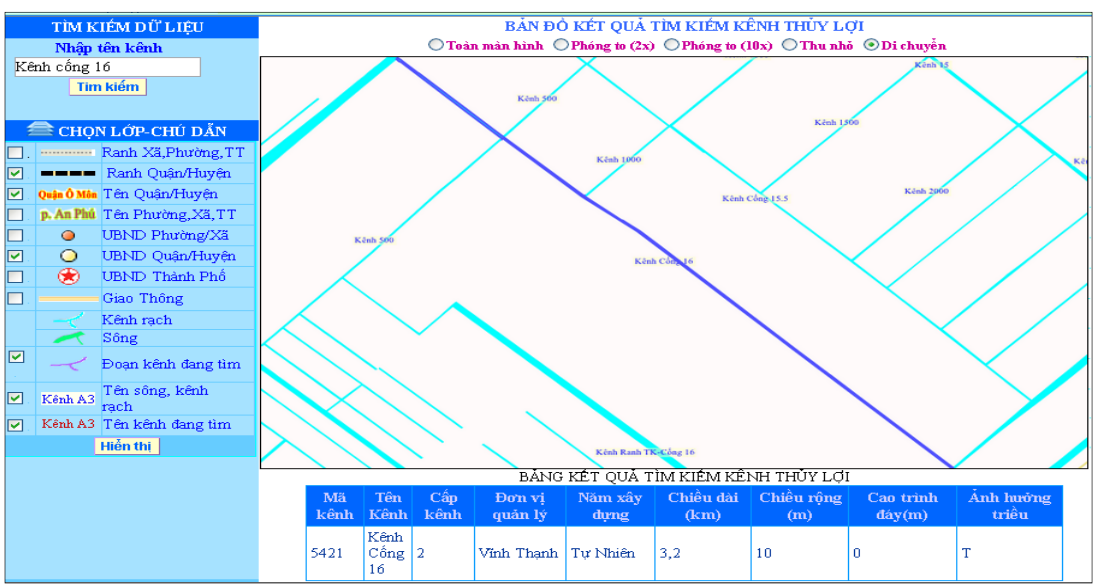
Dựa vào thư viện lập trình GIS nguồn mở SharpMap cho phép AGRIWEBGIS_CT tích hợp được các công cụ thao tác trên các lớp bản đồ như: Thanh công cụ hỗ trợ xem bản đồ (Xem toàn màn hình, phóng to, thu nhỏ và di chuyển), điều khiển chọn lớp hiển thị bản đồ giúp người dùng check vào ☑ để chọn lớp cần thể hiện nội dung bản đồ và kết xuất in bản đồ. Với lớp SharpMap. Rendering.Thematics và SharpMap.Styles được dịch ra từ thư viện SharpMap cho phép tô các màu phân biệt ở các vùng khác nhau và biểu diễn các đối tượng hình học dạng điểm và đường mô phỏng thế giới thực một cách sinh động. Hình 4 minh họa cách bố trí giao diện các trang bản đồ phân bố hiện trạng chăn nuôi – thú y của TP Cần Thơ trong đó tích hợp các lớp bản đồ GIS nền hành chính và các lớp chuyên đề về vị trí chăn nuôi heo, vịt, trâu, các cơ sở bán thức ăn, thuốc thú y và cơ sở thu gom chế biến trứng được điều tra thực tế lấy tọa độ GPS.



Hình 4: Trang bản đồ phân bố hiện trạng chăn nuôi, thú y thuộc phân hệ Quản lý dữ liệu Thú y

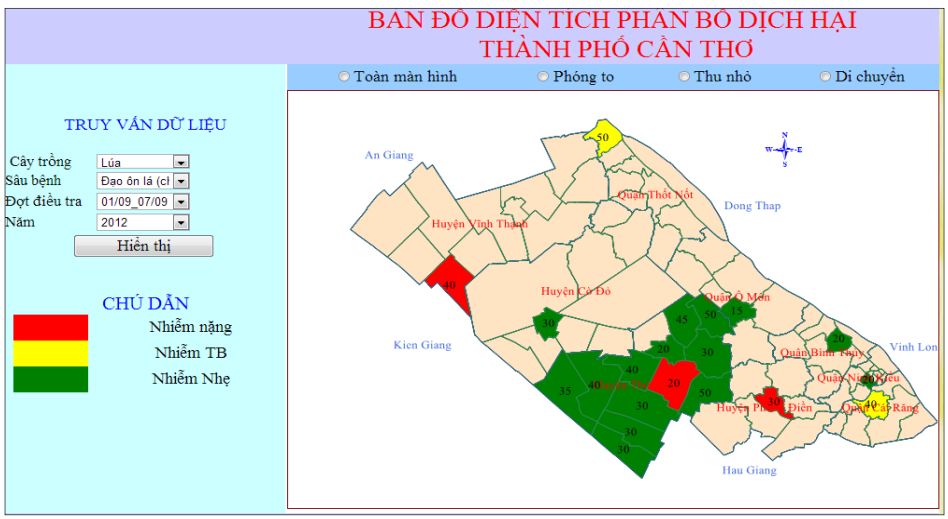
3.3.2 Chức năng tìm kiếm dữ liệu trên bản đồ

Một trong những đặc trưng quan trọng của GIS là xác định vị trí, nó cho phép truy cập và xác định "Cái gì ? ở đâu ?". Đặc trưng này được áp dụng khá linh hoạt vào các phân hệ của AGRIWEBGIS_CT. Thông qua lớp thư viện SharpMap.Data giúp thực hiện phương pháp ExecuteIntersectionQuery để truy xuất dữ liệu bản đồ song song với thủ tục kết nối ADO.NET để thể hiện bảng dữ liệu thuộc tính. Nội dung chung các trang tìm kiếm này là nhập tên đối tượng hoặc chọn điều kiện để truy vấn. Kết quả truy vấn dữ liệu sẽ được trình bày dưới biểu mẫu dạng bảng và trên bản đồ. Trường hợp, trong phân hệ quản lý dữ liệu Thủy lợi, để biết vị trí một tuyến kênh thủy lợi trên bản đồ cùng với các số liệu về công trình kênh đó khi cấp quản lý có nhu cầu nạo vét, tu bổ lại tuyến kênh. Người dùng nhập tên kênh cần tra cứu và bấm nút "Tìm kiếm dữ liệu", kết quả cho người dùng về hình dạng tuyến kênh đó được tô màu xanh dương làm dấu trên bản đồ kênh rạch trực tuyến (Hình 5).



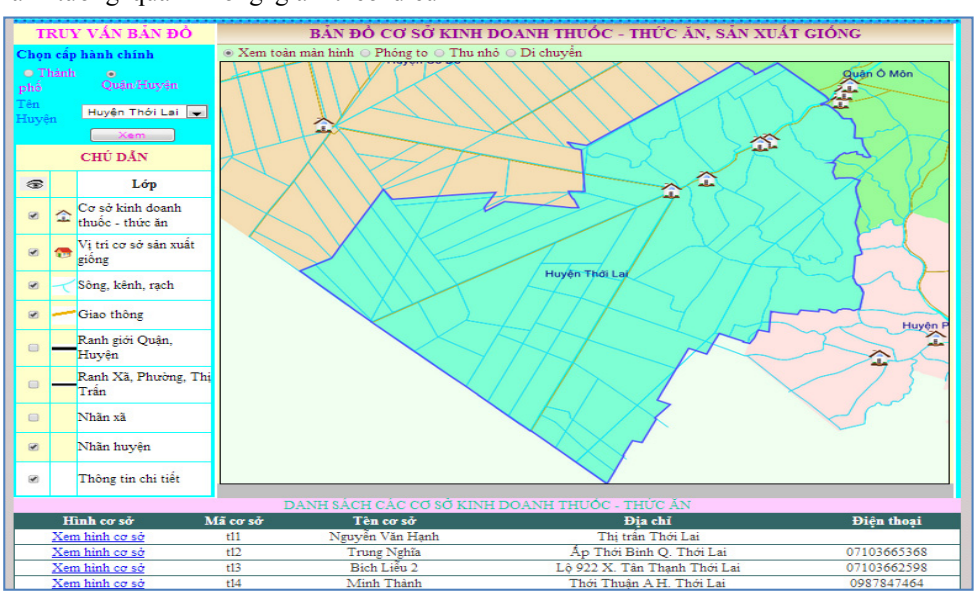
Hình 5: Trang tìm kiếm thông tin kênh theo tên thuộc phân hệ Thủy lợi

Các chức năng tìm kiếm kết hợp dữ liệu thuộc tính và không gian để tạo bản đồ chuyên đề động thể hiện qua trường hợp Tạo bản đồ phân bố dịch hại trong phân hệ quản lý dữ liệu bảo vệ thực vật (Hình 6), cho phép chọn nhiều điều kiện để truy vấn với phương pháp sử dụng chồng lớp bản đồ Intersect được hỗ trợ cho xử lý dữ liệu không gian trên SQL Server 2008 giữa lớp bản đồ nền hành chính và các số liệu điều tra dịch hại, xuống giống và thu hoạch để hiển thị kết quả trên bản đồ. Trang bản đồ phân bố dịch hại với các điều kiện truy vấn như: chọn loại cây trồng, loại sâu bệnh, đợt điều tra và năm sẽ biểu diễn được phân bố các cấp độ nhiễm dịch hại trên lúa ở các xã và diện tích nhiễm. Từ kết quả này, Chi cục Bảo vệ thực vật có thể khoanh vùng đưa ra khuyến cáo phòng trừ kịp thời.



Hình 6: Trang tìm kiếm phân bố tình hình dịch hại theo thuộc phân hệ BVTV

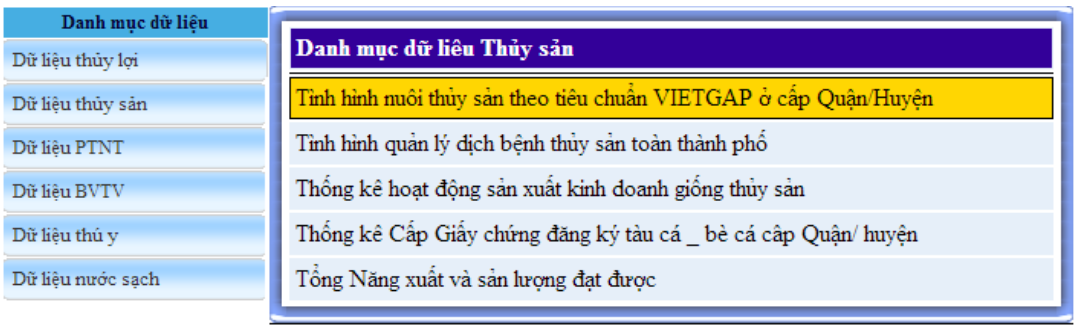
Chức năng tìm kiếm dạng chọn đối tượng theo tương quan không gian giữa 2 lớp thể hiện trong phân hệ quản lý dữ liệu Thủy sản với chức năng tìm các cơ sở kinh doanh thuốc, thức ăn và sản xuất giống thủy sản. Trong đó hệ thống thực hiện chọn với hàm tương quan không gian theo điều kiện các cơ sở giao (Intersect) với đối tượng lớp hành chính huyện. Hình 7 thể hiện kết quả tìm các điểm kinh doanh thuốc và thức ăn trong địa bàn huyện Thới Lai và hiển thị kết quả thuộc tính của từng điểm có hình ảnh của cơ sở kèm theo.



Hình 7: Tìm cơ sở kinh doanh thuốc, thức ăn, giống thủy sản theo đơn vị hành chính

3.3.3 Chức năng thống kê số liệu

Các số liệu thuộc tính thu thập được từ điều tra thực tế ở các chi cục được lưu trữ vào CSDL của phân hệ và được truy xuất báo cáo thống kê từ phân hệ tổng hợp của Sở hoặc từ các phân hệ quản lý (Hình 8). Kết quả thống kê sẽ tạo ra các bảng báo cáo số liệu theo nội dung chuyên đề của từng chuyên đề. Điều này giúp tiết kiệm thời gian và công sức tổng hợp báo cáo theo hình thức thủ công truyền thống. Hình 9 thể hiện một kết quả tổng hợp thông tin đường ống cấp nước ở trong phân hệ quản lý nước sạch. Kết quả báo cáo có thể được xuất sang định dạng báo biểu thiết kế bằng công cụ Crystal Report theo quy định của đơn vị quản lý hoặc xuất sang các định dạng khác như PDF, Excel.



Hình 8: Truy xuất dữ liệu tổng hợp từ phân hệ tổng hợp

Thông Tin Tổng Hợp Chiều Dài Đường Ống Cấp Nước

Chọn cấp hành chính: Thành phố Quận/Huyện Xã/Phường/Thị trấn

Chọn tên Quận/ Huyện: Quận Cái Răng

Chọn tên Xã/Phường/Thị trấn: p. Ba Láng

Xem Thông Tin | Thống kê chiều dài đường ống cấp nước | Xuất Báo Cáo

Thông Tin Chiều Dài Đường Ống

Mã Mạng Lưới	Ấp/Khu vực	Đường Ống Tổng (m)	Chiều Dài Nhánh 1 (m)	Chiều Dài Nhánh 2 (m)
CR001	KV 6	1990	850	1100
CR002	KV 3	1967	983	964
CR003	KV 6	1217	552	610
CR004	KV 4	1876	990	850
CR005	Phú Lễ	2919	627	1532
CR006	Phú Thuận A	2195	685	690
CR007	Khu vực 5	2692	1040	1652
CR008	Trạm tập trung	1426		
CR009	Thạnh Phú	1917	840	1127
CR010	An Hưng	1714	835	853
CR011	Phú Hưng	2054	885	1134
CR012	Khánh Hưng	2001	1147	844
CR013	Khánh Bình	1997	1003	979
CR014	Trạm tập trung	7494		
CR015	Thạnh Mỹ	1998	1091	882

1 2

Hình 9: Chức năng thống kê thông tin tổng hợp thuộc phân hệ Nước sạch

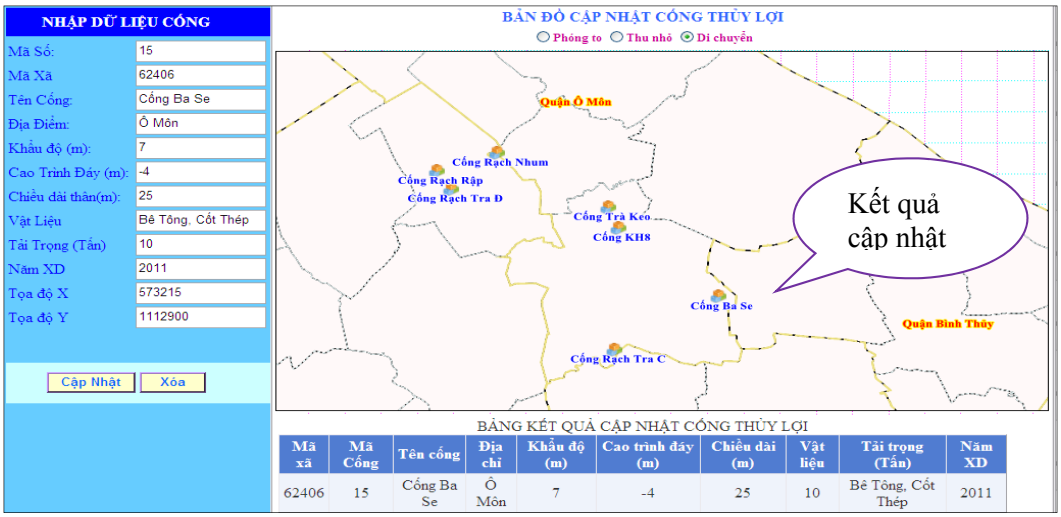
3.3.4 Chức năng quản trị hệ thống

AGRIWEBGIS_CT được thiết kế với các nhóm chức năng chính để quản trị hệ thống:

– Quản lý người dùng: Cung cấp tính năng quản lý thông tin và thêm người dùng.

– Nhập liệu: Nhằm giúp người quản trị hệ thống các phân hệ thực hiện công việc quản lý dễ dàng, chỉnh sửa dữ liệu được thực hiện nhanh chóng trên trình duyệt web không cần mở nguồn CSDL gốc. Tại đây, người quản trị có thể thực hiện các thao tác cập nhật, chỉnh sửa và xoá dữ liệu trong CSDL của hệ thống website. Về cập nhật dữ liệu không gian của hệ thống. Các phân hệ hỗ trợ cập nhật trực tiếp trên web cho dữ liệu không gian vector dạng điểm thông qua Microsoft.SqlServer. Types được hỗ trợ trong SQL SERVER 2008 cùng với quá trình cập nhật dữ liệu thuộc tính. Hình 10 thể hiện kết quả cập nhật dữ liệu cống thủy lợi với tọa độ thu được từ GPS, hệ thống sẽ vẽ thêm điểm trên bản đồ bằng toạ độ x, y theo phép chiếu UTM kết hợp lưu trữ dữ liệu thuộc tính của đối tượng đã nhập.

Đối với dữ liệu vector dạng đường và vùng, hệ thống không hỗ trợ cập nhật trực tiếp trên Web, người quản lý hệ thống sử dụng phần mềm Quantum GIS để kết nối và cập nhật bản đồ trực tiếp trên CSDL không gian.



Hình 10: Giao diện trang cập nhật cống thuỷ lợi mới

3.4 Đánh giá kết quả thực hiện

Hệ thống AGRIWEBGIS_CT trải qua 3 giai đoạn thực hiện chính như sau:

– Giai đoạn từ tháng 5/2011 đến 4/2012: thu thập dữ liệu thứ cấp tại các chi cục, trung tâm thuộc Sở và phối hợp điều tra thực tế để thu thập bổ sung nguồn số liệu.

– Giai đoạn từ tháng 4/2012 đến 5/2013: chuẩn hóa và phân tích dữ liệu cho hệ thống. Tiến hành lập trình xây dựng các ứng dụng cho phân hệ quản lý trên nền web.

– Giai đoạn từ tháng 5/2013 đến 12/2013: cài đặt thử nghiệm hệ thống vào Server và triển khai ứng dụng thực tế cho Sở quản lý tại địa chỉ:www.gissonongngnhiepcantho.gov.vn. Hiện tại, hệ thống được vận hành thử nghiệm đã đáp ứng các yêu cầu trong công tác quản lý dữ liệu của Chi cục, Sở với việc quản lý số liệu, bản đồ tích hợp công nghệ GIS trực tuyến một cách thuận tiện hơn, nguồn dữ liệu thống nhất và dễ dàng truy xuất các báo cáo ở mọi nơi và tiết kiệm thời gian.

Hệ thống Website được đánh giá về mặt xã hội sẽ cung cấp nhiều thông tin bổ ích cho người dân tại địa phương và các tỉnh thành khác về tình hình sản xuất, phát triển của lĩnh vực nông nghiệp và phát triển nông nghiệp của Thành phố Cần Thơ.

Với chuẩn dữ liệu không gian lưu trên hệ quản trị này có thể rút trích ra các định dạng dữ liệu không gian như: ESRI Shapefile, AutoCAD DXF, MapInfo file, Microstation DGN, Geography Markup Language GML, Keyhole Markup Language KML để kết nối và chia sẻ dữ liệu vào các ứng dụng GIS khác trong và ngoài thành phố.

4 KẾT LUẬN

Hệ thống AGRIWEBGIS_CT được thiết kế hoàn chỉnh và từng bước triển khai ứng dụng thực tế với một phân hệ quản lý tổng hợp và 6 phân hệ con cho 6 đơn vị trực thuộc Sở. Mỗi phân hệ cung cấp các bản đồ chuyên đề trực tuyến và chức năng quản lý dữ liệu phục vụ cho công tác quản lý hiện trạng ở 6 lĩnh vực trực thuộc Sở Nông nghiệp và Phát triển Nông thôn Thành phố Cần Thơ. Việc sử dụng thư viện lập trình SharpMap kết hợp xây dựng dữ liệu không gian và dữ liệu thuộc tính trên hệ quản trị cơ sở dữ liệu quan hệ SQL SERVER 2008 cho phép xây dựng một hệ thống WebGIS quản lý dữ liệu hiệu quả và nhanh chóng, chuẩn dữ liệu được nhiều phần mềm GIS hỗ trợ, dễ dàng chuyển định dạng dữ liệu trong quá trình khai thác.

LỜI CẢM TẠ

Nhóm tác giả xin chân thành cảm ơn Sở Khoa học và Công nghệ Thành phố Cần Thơ - đơn vị tài trợ kinh phí thông qua Dự án "Ứng dụng kỹ thuật GIS trong quản lý phát triển nông nghiệp - nông thôn".

Xin cảm ơn các đơn vị đã tham gia cung cấp dữ liệu và góp ý cho dự án.

TÀI LIỆU THAM KHẢO

1. B.Veeranna, I.V.Muralikrishna & E.G. Rajan, (2007). WebGIS – an application of agriculture information system at district level, Journal of Geomatics, 1(2), 002
2. Ju, X., & Chen, Z. (2011). Design and Implementation of WebGIS Based on

.NET. Applied Informatics and Communication (Vol. 227, pp. 291-297): Springer Berlin Heidelberg.

3. Manoj Chandram Nair (2011). Spatial Data in Sql server, Microsoft Technology Center, Infosys Technologies. Michelle Dumler, 2007.
4. Peter DeBetta, Greg Low and Mark Whitehorn, (2008). Introducing Microsoft SQL Server 2008, Microsoft Press, Microsoft Corporation.
5. VIETRADE – Cục xúc tiến thương mại (2012). Tình hình phát triển kinh tế tại Cần Thơ. Truy cập từ www.vietrade.gov.vn, ngày 30/08/2013.
6. Zhang, H., Xi, L., Ma, X., Lu, Z., Ji, Y., & Ren, Y. (2008). Research and Development of the Information Management System of Agricultural Science and Technology to Farmer Based on GIS. Computer And Computing Technologies In Agriculture, Volume I (Vol. 258, pp. 141-150): Springer US.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

ĐIỀU KHIỂN ROBOT PIONEER P3-DX BÁM SÁT ĐỐI TƯỢNG

Mã Trường Thành[1], Phạm Nguyên Khang[2], Lâm Chí Nguyện[3], Triệu Thanh Ngoan[3], Hà Duy An[3] và Đỗ Thanh Nghị[3]

[1] *Khoa Kỹ thuật - Công nghệ, Trường Cao đẳng Cộng đồng Sóc Trăng*
[2] *Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ*
[3] *Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Control of robot Pioneer P3-DX for object tracking

Từ khóa:
Camshift, meanshift, phát hiện đối tượng, bám sát đối tượng, đặc trưng haar-like, adaboost, robot Pioneer, cấu trúc phân tầng

Keywords:
Camshift, meanshift, object detection, tracking object, haar-like feature, adaboost, robot Pioneer, cascades of boosted classifers

ABSTRACT

In this paper, we present how to combine the machine learning algorithms into Pioneer P3-DX Robot that control the Robot's behaviors, movements, and tracking according to the real time. The machine learning algorithm, Cascades of Boosted Classifiers with Haar-Like Feature, can be used to detect and regconize objects. Next, we used Camshift Algorithms (Continiously Adaptive Meanshift) to monitor the object's activities within a flat image. Finally, based on the object's size and position in the flat image, we control the Robot to move and keep track of the object. The experimental results so that the Robot has been operated as the target objectives such as controlling behaviors, moving and keeping track of an object.

TÓM TẮT

Trong bài này, chúng tôi trình bày cách kết hợp các giải thuật máy học vào Robot Pioneer P3-DX nhằm điều khiển các hành vi, di chuyển và bám sát của Robot theo thời gian thực. Giải thuật máy học Cascades of Boosted Classifiers sử dụng đặc trưng Haar-Like Feature có thể phát hiện, nhận dạng đối tượng. Tiếp theo chúng tôi sử dụng giải thuật CamShift (Continuously Adaptive MeanShift) để giám sát hoạt động của đối tượng trong phạm vi mặt phẳng ảnh. Bước cuối cùng là tính toán vị trí và kích thước của đối tượng trên mặt phẳng ảnh để điều khiển Robot hoạt động và bám sát theo đối tượng. Kết quả thử nghiệm cho thấy Robot đã hoạt động đảm bảo các mục tiêu đặt ra như điều khiển các hành vi, di chuyển và bám sát đối tượng.

1 GIỚI THIỆU

Nghiên cứu điều khiển Robot là bài toán được các nhà khoa học quan tâm, nhằm phục vụ cho con người ở nhiều lĩnh vực ứng dụng như: Robot khuân hàng hóa theo chủ của nó, Robot dọn rác, lau nhà, hỗ trợ người khuyết tật, Robot trong công nghiệp ôtô, thăm dò khai thác mỏ, ... Do có tính ứng dụng cao nên các nhà nghiên cứu đã bắt tay vào phát triển Robot thông minh hơn, phục vụ tốt cho nhu cầu phát triển kinh tế, xã hội. Để làm được điều đó, cần có sự kết nối giữa "bộ não thông minh" và Robot để tạo nên những Robot thông minh như ngày nay. Máy học chính là nền tảng giúp Robot có thể thông minh, hoạt động tinh vi hơn. Trong các hướng nghiên cứu khác nhau về Robot, nghiên cứu theo dõi, bám sát đối tượng là một trong những bài toán hay và được ứng dụng nhiều trong thực tế. Những Robot này có thể thực hiện nhiều hành động khác nhau thông qua hành vi của đối tượng

để đáp ứng tốt cho con người nhờ vào sử dụng các mô hình máy học tự động. Hiện nay các nghiên cứu liên quan đến phát hiện và bám sát đối tượng bao gồm các hướng tiếp cận như sau: dự đoán chuyển động của đối tượng đang quan sát như bộ lọc Kalman [7][5] kết hợp với thuật toán MeanShift [14]do Fukunaga và Hostetler trình bày nhằm mục đích nâng cao khả năng việc bám sát. Hướng tiếp cận bằng phương pháp xây dựng mô hình cũng được nhiều tác giả thực hiện, chẳng hạn như Anurag Mital và Mikos Paragios [3] đã dùng phương pháp ước lượng mật độ bằng hàm nhân thích nghi (*adaptive kernel density estimation*) để tính toán. Phương pháp này được đánh giá là tốt về mặt không gian lưu trữ nhưng về việc tính toán trên thời gian thực thì còn hạn chế bởi tính toán phức tạp và tốc độ chậm. Ngoài ra, hướng tiếp cận được nhiều người hướng đến là bám sát theo màu sắc như thuật toán CamShift [11], MeanShift [20] hay hướng tiếp cận theo đặc trưng góc cạnh của Cox và Hingorani đã sử dụng MHT [12] (*Multiple Hypothesis Tracking*), hoặc là sự kết hợp của "particle filters" và kỹ thuật tính toán xác suất do MacCormick và Blake [12]. Nhìn chung, các cách tiếp cận đều có những ưu nhược điểm riêng nhằm giải quyết các bài toán khác nhau về giám sát. Với những hướng tiếp cận trên đều tiến tới mục đích là làm cho Robot di chuyển và bám theo đối tượng. Hầu hết các công trình nghiên cứu đều thực hiện về bám theo người [6][9] hay bám sát theo quả bóng [1]và những vật thể khác. Nếu nói về các thuật toán dùng để bám đối tượng trong ảnh rất đa dạng, phải kể đến một số thuật toán như: Thuật toán KLT, Correlation-base template matching, thuật toán Meanshift, thuật toán Camshift.

Trong bài báo này, chúng tôi áp dụng thuật toán Camshift để bám sát đối tượng dựa trên màu sắc đối tượng trên ảnh để Robot di chuyển theo, do tính hiệu quả về độ chính xác, độ phức tạp về thời gian. Chúng tôi đã sử dụng Robot Pioneer P3-DX kết hợp với máy học để giải quyết bài toán: "Điều khiển Robot bám sát đối tượng". Chúng tôi trích các đặc trưng Haar-Like [18][16] kết hợp với bộ phân tầng Cascades of Boosted Classifiers và Adaboost [16][17] để huấn luyện cho Robot nhận dạng đối tượng - đối tượng được sử dụng để Robot nhận dạng là quả bóng. Sau khi Robot phát hiện và nhận dạng được đối tượng, chúng tôi sẽ sử dụng giải thuật CamShift để Robot bám lấy quả bóng. Từ đó, dựa trên mặt phẳng ảnh có đối tượng đang hiện hữu trong đó sẽ cho Robot di chuyển thông qua việc tính khoảng cách và lấy tỷ lệ cho vận tốc để Robot di chuyển theo đối tượng quan sát. Nhờ vào việc xám hóa ảnh khi phát hiện và nhận dạng sẽ giúp cho đối tượng có thể linh động, giải thuật CamShift [11] sẽ hiệu quả hơn trong bám sát. Với trường hợp, giải thuật CamShift bám sai đối tượng thì bộ phận phát hiện và nhận dạng đối tượng sẽ liên tục tìm đối tượng trên ảnh *(lúc này Robot sẽ đứng yên hoặc xoay tìm đối tượng)* cho đến khi phát hiện và giải thuật CamShift được thực hiện. Kết quả cho thấy Robot đã hoạt động như mục tiêu đặt ra và hạn chế được tình trạng mất bám trong các môi trường khác nhau cũng như kiểm soát được khoảng cách và vận tốc di chuyển của Robot.

Phần tiếp theo của bài báo được tổ chức như sau: phần 2 trình bày ngắn gọn về Robot di động – Robot Pioneer P3-DX. Phần 3: trình bày điều khiển Robot bám sát đối tượng: phát hiện và nhận dạng đối tượng với Haar-Like feature kết hợp Cascades of Boosted Classifiers và Adaboost, trình bày thuật toán Camshift để bám sát đối tượng và cách di chuyển của Robot trên mặt phẳng ảnh, cách tính khoảng cách tương ứng với vận tốc điều khiển Robot. Kết quả thực nghiệm được trình bày ở phần 4 trước khi kết luận và hướng phát triển trong phần 5.

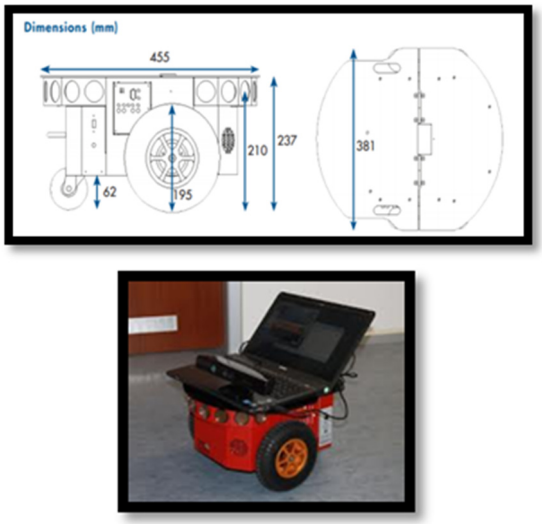
2 ROBOT DI ĐỘNG – ROBOT PIONEER P3-DX

Mobile Robot hay còn gọi là Robot di động, Robot tự hành. Đây là một loại Robot có vai trò quan trọng trong ngành Robot học. Cùng với sự phát triển của các hệ thống về lĩnh vực công nghệ thông tin, Robot di động đã góp phần mạnh mẽ trong công nghiệp, y tế và cuộc sống thực tế hiện nay. Robot di động là một loại máy móc tự động có khả năng di chuyển trong một phạm vi môi trường nhất định.

Robot được sử dụng trong bài viết là Robot Pioneer P3-DX [2]. Đây là dòng Robot di động chế tạo bởi hăng Adept Mobile Robot. Robot Pioneer là một trong những Robot được nhiều nhà chuyên gia đa lĩnh vực nghiên cứu đến và được sử dụng phổ biến nhất trên thế giới hiện nay. Robot này được tạo ra và cho phép người nghiên cứu hoàn toàn có thể ***"lập trình được"***.

Robot Pioneer P3-DX [2]là loại Robot di động nhỏ, trọng lượng nhẹ với 3 bánh xe (2 bánh trước chủ động *(bánh chính)* và bánh sau di chuyển tự do *(bánh phụ - rẽ)*), bánh xe với đường kính 19.5 cm, thân được bao bọc bởi kim loại nhôm cứng cáp. Phía trước được trang bị 8 cảm biến Sonar, phía sau cũng được trang bị 8 cảm biến Sonar – đây là dạng cảm biến siêu âm để dò tìm vật cản. Robot

Pioneer P3-DX có 3 pin nhằm dự trữ và có thể thay đổi nhanh chóng.



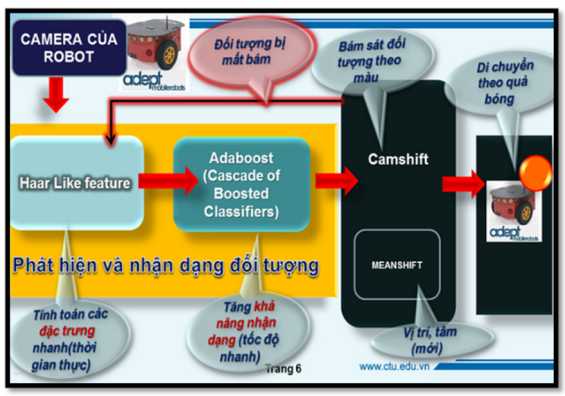
Hình 1: Cấu tạo Robot Pioneer P3-DX [2]

Vì khả năng có thể lập trình được nên hãng đã hỗ trợ bộ SDK *(software development kit)* dành cho những nhà nghiên cứu Robot, các loại máy tính có thể dễ dàng cài đặt và tích hợp vào để Robot có thể hoạt động nhịp nhàng.

Robot Pioneer P3-DX có thể di chuyển tới và di chuyển lui với vận tốc tối đa là 1.2 m/s và có tải trọng đồ vật lên đến 23 kg.

3 ĐIỀU KHIỂN ROBOT PIONEER P3-DX BÁM SÁT ĐỐI TƯỢNG

Hệ thống đề xuất cho Robot bám sát đối tượng



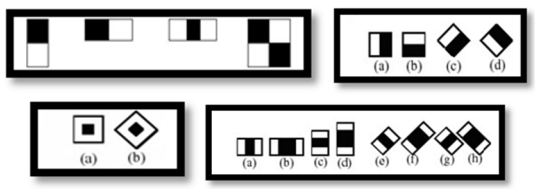
Hình 2: Mô hình hoạt động của hệ thống

Robot sẽ được tích hợp một Webcam/Camera để quan sát, dựa vào frame ảnh thu được hệ thống sẽ sử dụng Haar-Like features và Adaboost để phát hiện đối tượng cần nhận dạng trên từng frame ảnh. Dựa vào đối tượng đã được nhận dạng ở trên, hệ thống sẽ sử dụng thuật toán CamShift để bắt bám đối tượng trong quá trình di chuyển *(sử dụng MeanShift để lấy vị trí và trọng tâm mới của đối tượng)*. Sau khi đã bắt, bám được đối tượng thì hệ thống sẽ dựa vào kích thước và mặt phẳng ảnh để điều khiển Robot di chuyển. Trong trường hợp thuật toán CamShift bị mất bám thì bộ phận phát hiện và nhận dạng đối tượng sẽ thực hiện lại việc dò tìm.

3.1 Phát hiện và nhận dạng đối tượng

3.1.1 Trích đặc trưng Haar-Like và Cascades of Boosted Classifiers

Đặc trưng Haar-Like [18][16]là những đặc trưng ảnh số được sử dụng trong nhận dạng đối tượng. Thông thường, những đặc trưng cơ bản Haar-Like dùng để dò khuôn mặt trong thời gian thực, do đó Haar-Like là sự lựa chọn phù hợp cho hệ bám sát đối tượng nhằm nhận dạng nhanh chóng, chính xác trong thời gian thực.

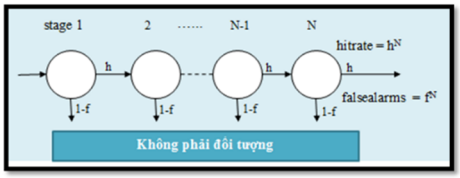


Hình 3: Đặc trưng Haar-Like của Viola và Jones [16]

Đặc trưng Haar-like sẽ diễn đạt được tri thức về các đối tượng trong ảnh *(bởi vì nó biểu diễn mối liên hệ giữa các bộ phận của đối tượng)*, điều mà bản thân từng điểm ảnh không diễn đạt được – dựa trên việc co giãn của khung cửa sổ tìm kiếm. Để tính giá trị các đặc trưng Haar-Like, cần tính sự chênh lệch giữa tổng của các pixel của các vùng đen và các vùng trắng như trong công thức sau:

$$f(x) = \sum_{đen}(pixel) - \sum_{trắng}(pixel) \qquad (1)$$

Lý do đặc trưng Haar-Like được tính toán nhanh chính là nhờ vào khái niệm tích phân ảnh *(integration)* mà Viola và Jone đã đề xuất. Kết hợp với bộ cấu trúc phân tầng với sự huấn luyện của một tầng *(cascade)* gồm *Ngiai đoạn (stages)*. Ở mỗi giai đoạn, phân lớp yếu *(weak classifier)* tương ứng sẽ được huấn luyện sao cho độ chính xác của nó là *h (tỷ lệ nhận dạng chính xác đối tượng)* và false alarm bằng *f (tỷ lệ nhận dạng lầm)*.



Hình 4:Cấu trúc phân tầng với N giai đoạn[16]

3.1.2 Adaboost

AdaBoost [19] (*Adaptive Boost*) được Freund và Schapire đề xuất trong những năm 1995 là một bộ phân loại mạnh phi tuyến. Adaboost hoạt động trên nguyên tắc kết hợp tuyến tính các phân lớp yếu *(weak classifier)*. Nghiên cứu của Viola và Jones [16] xây dựng Adaboost dựa trên dữ liệu đầu vào là các đặc trưng Haar- Like để hình thành một phân lớp mạnh hơn *(strong classifier)* theo nguyên tắc của boosting [19]: ***"tăng trọng số của các mẫu bị nhận dạng sai và giảm trọng số của các mẫu được nhận dạng đúng bởi phân lớp yếu*** *(weak classifier)* ***vừa xây dựng"***. Ý tưởng chính của Boosting là thực hiện lặp lại nhiều lần quá trình học của bộ phân lớp yếu, nghĩa là tập trung học lại các phần tử bị phân lớp sai. Tại mỗi giai đoạn (stages) của Cascades of Boosted Classifiers được huấn luyện là một Adaboost.

Các bộ phân lớp yếu $h_k(x)$ là các bộ phân lớp yếu được lấy từ việc trích các đặc trưng Haar-Like, được biểu diễn như sau [16]:

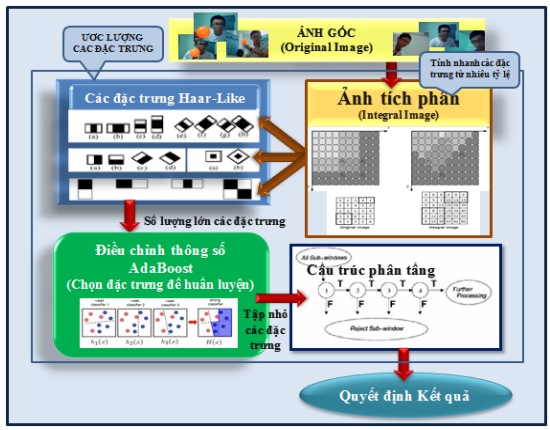
$$h_k(x) = \begin{cases} 1 & \text{nếu } p_k f_k(x) < p_k \theta_k \\ 0 & \text{nếu ngược lại} \end{cases} \quad (2)$$

- X : mẫu hay cửa sổ con cần xét *($X = (x_1, x_2, ..., x_n)$ là vector đặc trưng mẫu);*
- θ_k: ngưỡng (θ = teta);
- f_k: giá trị của đặc trưng Haar-like (1);
- p_k: hệ số chiều của bất phương trình *(a parity indicating the direction of the inequality sign).*

Bộ phân lớp mạnh $h^{mạnh}(x)$ được tạo ra bằng cách kết hợp tuyến tính N bộ phân loại yếu (2). Vấn đề cốt lõi là thực hiện chọn các trọng số $\alpha_1, \ldots, \alpha_n$ thích hợp [19].

$$h^{mạnh}(x) = sign \sum_{k=1}^{N} \alpha_k . h_k^{yếu}(x) \quad (3)$$

3.1.3 Nhận dạng và phát hiện đối tượng



Hình 5: Mô hình phát hiện và nhận dạng đối tượng

3.2 Bám sát đối tượng với thuật toán Camshift

Thuật toán CamShift [11] (*Continuously Adaptive MeanShift*) được xuất phát và dựa trên nền tảng của thuật toán MeanShift, do vậy việc tính trọng tâm của đối tượng và dịch chuyển cửa sổ tìm kiếm sẽ thực hiện bằng MeanShift [14] [theo vết màu MeanShift - Comaniciu 2003] và Moment ảnh. Ý tưởng chính của Camshift là tính lại trọng tâm của đối tượng dựa vào frame ảnh trước đó và kích thước của cửa sổ tìm kiếm được thay đổi động theo đối tượng *(với MeanShift thì cửa sổ tìm kiếm là cố định).*

3.2.1 Thuật toán MeanShift [3] [20]

Bước 1: Khởi tạo vị trí của đích trong frame hiện tại với y_0, tính mô hình hiện tại đang xét (ứng viên) – $\{p(y_0)\}_{k=1..n}$

Tính:
$$\rho[p(y_o), q] = \sum_{k=0}^{n} \sqrt{p_k(y_0) q_u} \quad (4)$$

Bước 2: Tính được trọng số w_i theo

$$w_i = \sum_{k=0}^{n} \sqrt{\frac{q_k}{p_k(y_o)}} \delta(b(x_i) - k) \quad (5)$$

Bước 3: Tìm vị trí đích ứng viên kế tiếp theo

$$y_1 = \frac{\sum_i x_i w_i g\left(\left\|\frac{y_o - x_i}{h}\right\|^2\right)}{\sum_i w_i g\left(\left\|\frac{y_o - x_i}{h}\right\|^2\right)} \quad (6)$$

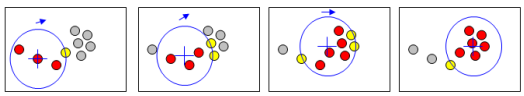
Bước 4: Tính ra mô hình mới $\{p(y_1)\}_{k=1_n}$ và tính được

$$\rho[p(y_1), q] = \sum_{k=0}^{n} \sqrt{p_k(y_1) q_u} \quad (7)$$

Bước 5: Khi (4) < (7) thì ta thực hiện:

$$y_1 = \frac{y_1 + y_0}{2} \quad (8)$$

Bước 5: Nếu $\|y_1 - y_0\| < \varepsilon$ thì cho thuật toán dừng, nếu không nhỏ hơn ta đặt $y_0 = y_1$ và cho quay lại bước 2 để tìm vị trí.



Hình 6: Kỹ thuật dịch chuyển cửa sổ tìm kiếm của MeanShift

3.2.2 Tính Moment ảnh

Những thuộc tính cơ bản của ảnh sẽ được tìm thấy thông qua Moment ảnh [15][11] bao gồm: vùng ảnh (area), trọng tâm của đối tượng và một số thông tin đi kèm với đối tượng đang xét. Đây là việc tính toán những đặc trưng Moment và ước lượng chiều dài, chiều rộng của đối tượng dựa vào Moment thứ 0.

Moment 2 chiều của bậc (i+j) của một ảnh f(x, y) có kích thước M x N sẽ được định nghĩa như sau [15]:

$$M_{ij} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} x^i y^j f(x, y) \quad (9)$$

Trong đó i, j là các số nguyên 0, 1, 2, 3,….

– Tính Moment thứ 0 [11]:

$$M_{00} = \sum_x \sum_y f(x, y) \quad (10)$$

– **T** ính Moment thứ nhất của x, y **Error! Reference source not found.**:

$$M_{10} = \sum_x \sum_y x f(x, y) \quad (11)$$

$$M_{01} = \sum_x \sum_y y f(x, y) \quad (12)$$

Một số tính chất cần quan tâm trong moment ảnh là:

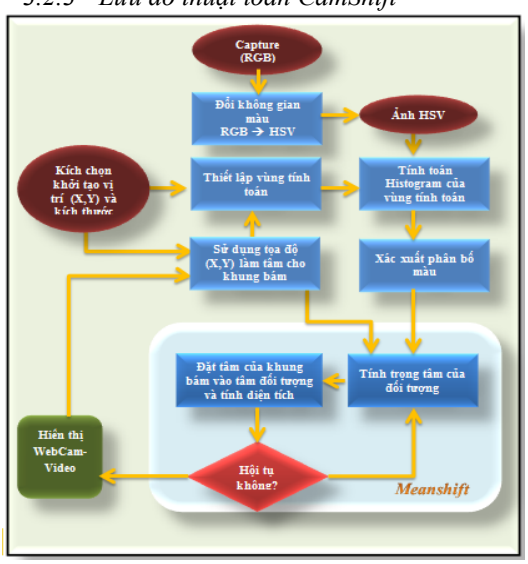
– Diện tích (cho ảnh nhị phân) hoặc tổng mức xám (ảnh xám) M_{00};

– Tâm của ảnh (đối tượng):

$$\{x_c, y_c\} = \left\{\frac{M_{10}}{M_{00}}, \frac{M_{01}}{M_{00}}\right\} \text{ REF So14}_{\text{MMercimekKGulezandTVMumc}} \text{\h * MERGEFORMAT} \quad (13)$$

[15] [11]

3.2.3 Lưu đồ thuật toán CamShift



Hình 7: Lưu đồ hoạt động của giải thuật CamShift [11]

3.3 Tính khoảng cách và di chuyển

3.3.1 Tính khoảng cách của đối tượng đến Camera

$$Z = f \times \frac{X}{x'} \rightarrow f = \frac{Z \cdot x'}{X} \quad (14)$$

*[Nguồn: **Error! Reference source not found.**-chapter 11.1]*

– Z: Khoảng cách vật đến Webcam;

– f: tiêu cự*(17,63-Webcam);*

– X: Kích thước vật (10 cm);

– x': Kích thước ảnh.

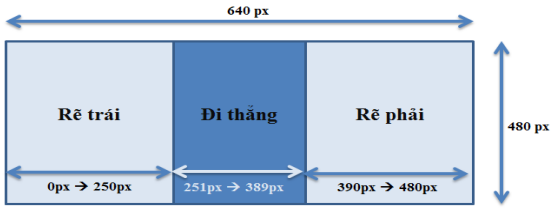
Với 1 pixel (X) = 0.0264583333333334 centimeter [cm].

$$Z\ (\text{Khoảng cách (cm)}) = \frac{17,63 \times X}{\text{Kích thước ảnh(cm)}} \quad (15)$$

$$\text{Vận tốc Robot} = \frac{V_đ \times Z}{100} \quad (16)$$

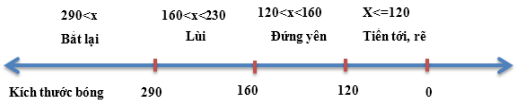
Với: V_d là vận tốc ban đầu.

3.3.2 Robot di chuyển theo mặt phẳng ảnh



Hình 8: Vị trí của quả bóng trên mặt phẳng ảnh ứng với hoạt động Robot

Robot chỉ thực hiện việc di chuyển khi quả bóng có kích thước nằm trong khoảng cho phép *(Hình 9)*. Khi kích thước ảnh của quả bóng quá lớn thì sẽ cho thực hiện bắt lại theo ngưỡng.



Hình 9: Kích thước của quả bóng ứng với hành vi của Robot

Nếu di chuyển qua vùng có cùng màu sắc với màu sắc mà thuật toán CamShift đang bắt thì hệ thống sẽ cho bắt lại *(vì mảng màu lớn x>290px)*.

4 KẾT QUẢ THỰC NGHIỆM

4.1 Mô tả thực nghiệm

– Bài báo nghiên cứu trên Robot Pioneer P3-DX của hãng Adept Mobile Robot được trình bày ở phần 2.

– Thông tin máy tính và ứng dụng cài đặt trong nghiên cứu:

Loại	Tên và kích thước sử dụng	Ghi chú
Hiệu máy tính	Laptop HP DV6T-2000;	
Bộ xử lý (CPU)	Intel Core i7 CPU Q720 @1.60Ghz (8 CPUs), 1.6GHz;	
Bộ nhớ (RAM)	4096MB – DDR3;	
Hệ điều hành	Ubuntu 12.04 OS (64 bit);	
Webcam	HP Webcam 1.3 MP;	
Tốc độ Webcam	20 frame/s;	
Thư viện cài đặt	OpenCV 2.4.2[10];*(gói thư viện cài đặt nhận dạng và bám sát)*	
Thư viện Robot	SDK ARIA 2.7.2 , MobileSim 0.5;	
Ngôn ngữ LT	C++, biên dịch g++ Linux	

4.2 Thiết lập thông số (giá trị) thực nghiệm

– Vận tốc ban đầu của Robot khi di chuyển là $V_0 = 200$ mm/s;

– Khoảng cách dừng của Robot trong khoảng $L_0 = (40; 55)$ cm;

– Khoảng cách Robot hoạt động: $L_t = 56$cm trở lên;

– Khoảng cách Robot lùi $L_b = (28; 40)$ cm;

– Khoảng cách xa nhất mà Robot không thể dò tìm đối tượng là: 333 cm (vì bộ huấn luyện nhận dạng với thông số min là 20 x 20 px → tương ứng với kích thước quả bóng);

– Khoảng cách ngắn nhất mà Robot không thể dò tìm đối tượng là: nhỏ hơn 22 cm*(290px);*

– Tốc độ ảnh xử lý được thiết lập: 20 frame/s;

– Tiêu cự Webcam được tính trung bình là: 17.6 cm;

– Kích thước giao diện Webcam: 640-480px.

4.3 Kết quả thực nghiệm

4.3.1 Huấn luyện nhận dạng và phát hiện đối tượng

– Thông số huấn luyện HaarTraining:

"-nstages 20 -nsplits 2 -minhitrate 0.999-maxfalsealarm 0.5 -npos 1600 -nneg 4200 -w 20 -h 20 -nonsym -mem 1100 -mode ALL"

- **-nstages 20**: Số lượng tầng của mô hình huấn luyện;
- **-nneg 4200:** Số lượng ảnh không có đối tượng (quả bóng) là 4200 ảnh;
- **-npos 1600:** Số lượng ảnh có đối tượng (quả bóng) là 1600 ảnh;
- **-nsplits:** Số lần phân hoạch dữ liệu là 2;
- **-mem 1100:** sử dụng 1.1 GB Ram *(bộ nhớ)* để training;
- **-minhitrate 0.999:** tỷ lệ nhận dạng đối tượng (quả bóng) tối thiểu là 0.999;
- **-maxfalsealarm 0.5:** tỷ lệ nhận dạng sai đối tượng tối đa là 50%;
- **-mode ALL:** sử dụng tất cả các đặc trưng bao gồm cả đặc trưng mở rộng;
- **-w 20:** kích thước chiều rộng; **-h 20:** kích thước chiều cao.

- Kết quả nhận dạng với độ chính xác *(recall)* **98%** khi kiểm thử với 200 ảnh có đối tượng quả bóng (Dương) và 200 ảnh không có đối tượng quả bóng (Âm):

Dự đoán =>	Dương	Âm
Dương	**196**	**4**
Âm	**1**	**199**

– Khi sử dụng công cụ **performance** trong OpenCV để ước lượng độ chính xác của mô hình huấn luyện, chúng tôi thu được độ chính xác *(recall)*là **97.4 %**.

4.3.2 Robot di chuyển

Dựa trên mặt phẳng ảnh *(mục 3.3.2)* và các thông số liên quan *(mục 4.2)* để điều khiển Robot, kết quả Robot đã di chuyển như sau *(phần phụ lục)*:

– Robot di chuyển thẳng tới *(Hình 10)*;

– Robot di chuyển rẽ phải *(Hình 11)*;

– Robot di chuyển rẽ trái *(Hình 12)*;

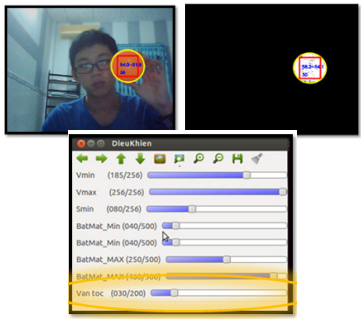
– Robot di chuyển lùi *(Hình 13)*.

5 KẾT LUẬN

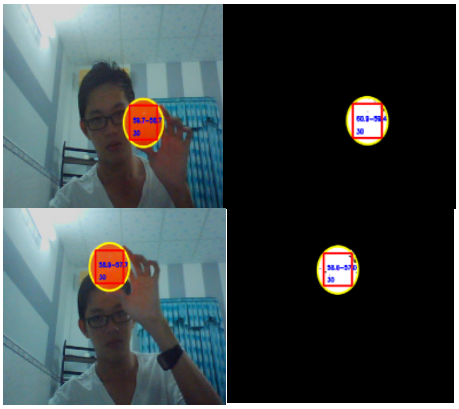
Chúng tôi vừa trình bày sự kết hợp giữa Robot Pioneer P3-DX và máy học để điều khiển các hành vi, di chuyển và bám sát của Robot theo thời gian thực. Robot có khả năng tự di chuyển đúng như yêu cầu đặt ra cho bài toán (*tiến tới, lùi, rẽ trái, rẽ phải*), giảm thiểu khả năng mất bám trong thời gian ngắn. Nhờ sự kết hợp của việc phát hiện đối tượng liên tục khi bị mất bám *(mất quan sát)* đã giúp Robot có thể nhanh chóng tìm lại đối tượng và di chuyển đúng yêu cầu. Robot di chuyển với vận tốc phụ thuộc vào khoảng cách từ đối tượng đến Robot. Tốc độ bắt bám đối tượng dao động 25÷30 ms.

Hướng nghiên cứu tiếp theo của bài báo là việc nâng cao khả năng bám sát đối tượng *(tracking object)* với việc tích hợp thêm bộ lọc Kalman để dự đoán đối tượng trên mặt phẳng ảnh. Robot trong nghiên cứu di chuyển động không theo một bản đồ cụ thể, do đó có thể tích hợp thêm một số bản đồ tĩnh nhằm giúp Robot có thể di chuyển trong một phạm vi nhất định và bám sát theo đối tượng *(ví dụ: Robot đá bóng)* hoặc Robot sẽ tránh được những chướng ngại vật trước mắt mà không làm ảnh hưởng khả năng bám sát.

Nghiên cứu các ứng dụng đáp ứng nhu cầu thực tế như: Robot chở hàng theo chủ nhân *(đối tượng nhận dạng là chủ nhân Robot)*, Robot làm phương tiện chuyên chở người khuyết tật *(đối tượng được người khuyết tật sử dụng để Robot di chuyển)*, Robot lau dọn vệ sinh *(đối tượng là các vết bẩn)*, Robot dập lửa trong khu vực nhỏ *(sử dụng kết hợp camera hồng ngoại và nhận dạng khu vực nhiệt độ cao trong ảnh hồng ngoại)*.



Hình 8: Bảng điều khiển và đối tượng được bám sát



Hình 9: Bắt bám theo đối tượng quả bóng

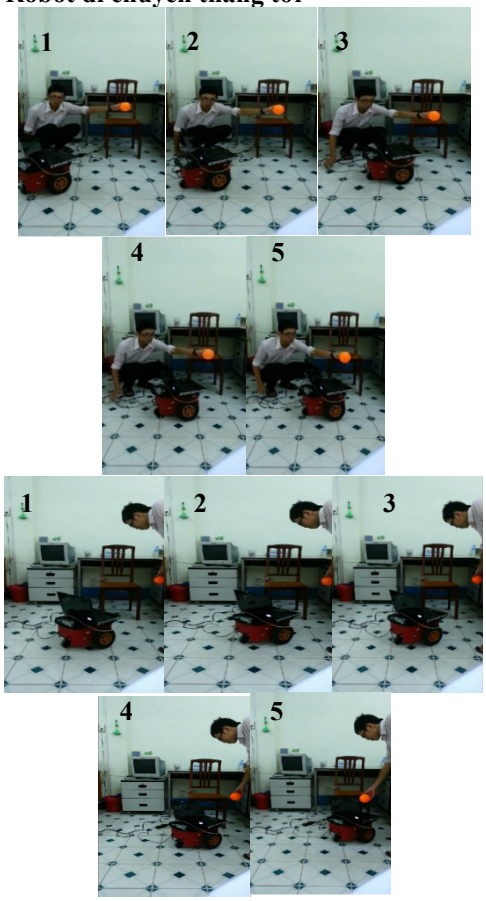
TÀI LIỆU THAM KHẢO

1. A. Treptow, A. Masselli and A. Zell (2003). *Real-Time Object Tracking for Soccer-Robots without Color Information*. University of Tuebingen Department of Computer Science, Gemany.
2. A. WhitBrook (2010). *Programming Mobile Robot with Aria and Player*. Springer.
3. A. Mittal and M. Paragios (2004). *Motion Based Background Subtraction using Adaptive Kernel Density Estimation*. pp. 302-309.
4. Artner (2008). *A Comparison of Mean Shift Tracking Methods*. Digital Media, Upper Austria University of Applied Sciences.
5. C. Zhang, Y. Qiao, E. Fallon, and C. Xu (2009). *An improved camshift algorithm for target tracking in video surveillance*. In Proceedings of 9th. IT and T Conference, number 12.
6. D. Beymer and Konolige K (2001). *Tracking people from a mobile platform*. In IJCAI-2001 Workshop on Reasoning with Uncertainty in Robotics.

7. D. Comaniciu and V. Ramesh (2009). *"Mean Shift and Optimal Prediction for Efficient Object Tracking,"* Proc. Int'l Conf. Image Processing, vol. III, pp. 70-73, 2000. 20 Sept.
8. D. Comaniciu, Dorin, Peter Meer (2002). *"Mean Shift: A Robust Approach Toward Feature Space Analysis"*. IEEE Transactions on Pattern Analysis and Machine Intelligence.
9. D. Kortenkamp, E. Huber, and R. P. Bonasso (1996). *Recognizing and interpreting gestures on a mobile robot*. InProc. of the American Conference on Artificial Intelligence.
10. G. Bradski and A. Kaehler (2012). *Learning OpenCV*. O'Reilly Media.
11. Gary R. Bradski (1998). *Computer Vision Face Tracking For Use in a Perceptual User Interface*. Microcomputer Reseacher Lab, Santa Clara, CA, Intel Corporation.
12. I.J. Cox and S.L (1996). Hingorani. *An efficient implementation of reids multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking*. IEEE Transactions on PAMI, 18(2):138–150.
13. J. MacCormick and A. Blake (1999). *A probabilistic exclusion principle for tracking multiple objects*. InProc. of 7th International Conference on Computer Vision (ICCV), pages 572–587.
14. K. Fukunaga and L.D. Hostetler (1975). *The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition*. IEEE Trans. Information Theory, vol. 21, pp. 32-40.
15. M. Mercimek, K. Gulez and T.V. Mumcu (2005). *Real object recognition using moment invariants*. Yildiz Technical University, Electrical-Electronics Faculty, Electrical Engineering Department, 34349 Besiktas-Istanbul, Turkey.
16. P. Viola and M. Jones (2001). *Robust Real-time Object Detection*. International Journal of Computer Vision.
17. Phạm Hồng Ngự (2009). *Nhận dạng đối tượng sử dụng thuật toán Adaboost*. Luận văn Thạc sỹ. Đại học Huế.
18. R. Lienhart, J. Maydt (2002). *An Extended Set of Haar-like Features for Rapid Object Detection*. Intel Labs, Intel Corporation, Santa Clara, CA 95052, USA.
19. Y. Freund, and R. Shapire (1995). *A decision-theoretic generalization of on-line learning and an application to boosting*. Proceedings of the Second European Conference on Computational Learning Theory. pp. 23-37.
20. Y.Cheng (1995). *Mean Shift, Mode Seeking, and Clustering*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, no. 8. August.

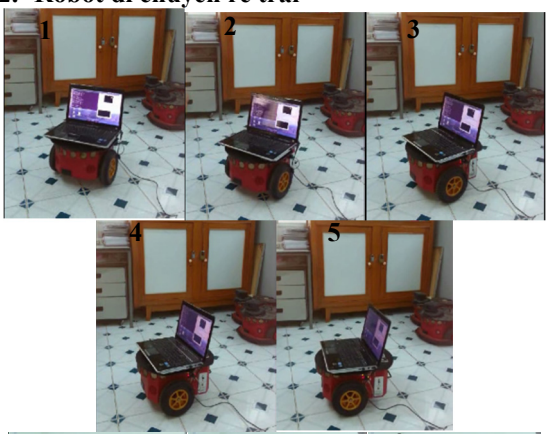
PHỤ LỤC

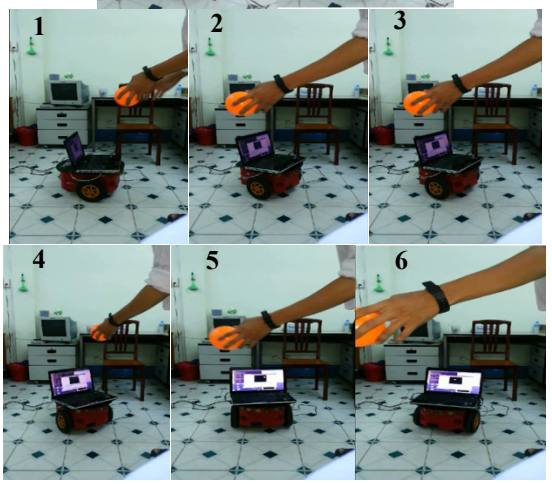
1. Robot di chuyển thẳng tới



Hình 10: Robot di chuyển thẳng tới

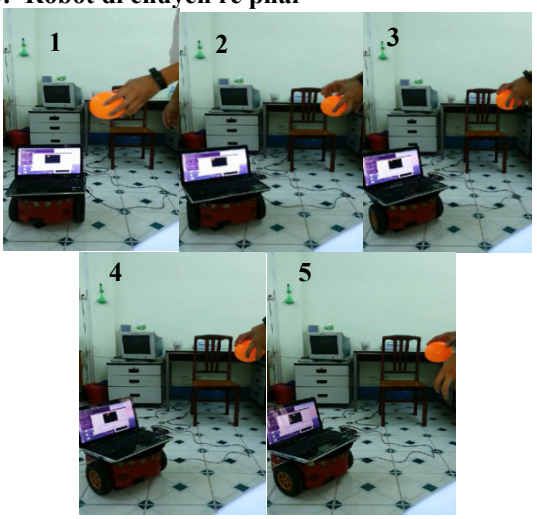
2. Robot di chuyển rẽ trái

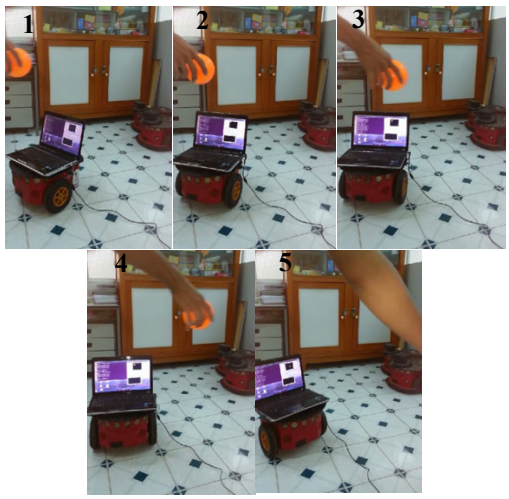




Hình 12: Robot di chuyển rẽ trái

3. Robot di chuyển rẽ phải





Hình 11: Robot di chuyển rẽ phải

4. Robot di chuyển lùi

Hình 13: Robot di chuyển lùi

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

GIẢI THUẬT SỬA LỖI CHO CÁC TRÌNH TỰ ADN NGẮN

Nguyễn Văn Hòa[1] và Nguyễn Văn Đông[1]
[1] *Khoa Công nghệ Kỹ thuật Môi trường, Trường Đại học An Giang*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Algorithm of short-read error correction

Từ khóa:
chuỗi ADN, xác lập trình tự ADN, chuỗi ADN ngắn, chỉ mục, sửa lỗi

Keywords:
DNA sequence, DNA sequencing, short read, kmer, error correction

ABSTRACT

Today with the development of DNA sequencing technology, we have obtained a large amount of DNA sequences in a short time with low cost. Specially, the next-generation DNA sequecing can generate a huge amount of short DNA sequences, called short reads with length from 30 to 100 bp. The short reads have an error rate between 1% and 2%. Therefore, the error reads must be corrected before being assembled into the complete genome. There are several proposed algorithms for correcting the error reads such as SHREC and SOAP de Novo. However, SHREC needs a long computation time to correct errors while SOAP de Novo requires very high memory usage. In this paper, we present our algorithm (RCorrector) based on the index structure of KMER for detecting and correcting error reads. Compared to the SHREC algorithm, the RCorrector algorithm provides a speed up from 3 to 7 with the same sensitivity and specificity.

TÓM TẮT

Ngày nay với sự tiến bộ của kỹ thuật xác lập trình tự ADN (DNA Sequencing) chúng ta có thể tạo ra một số lượng lớn các chuỗi ADN trong khoảng thời gian ngắn với chi phí thấp. Đặc biệt thế hệ xác lập trình tự mới hiện nay tạo ra số lượng rất lớn chuỗi ADN ngắn, được gọi là short read, với chiều dài từ 30 đến 100 nulcotide. Các read này có tỉ lệ lỗi từ 1% đến 2%. Do đó các read lỗi này phải được sửa lỗi trước khi được lắp ráp thành bộ gien ADN hoàn chỉnh. Nhiều giải thuật sửa lỗi đã được đề xuất như SHREC, SOAP de Novo. Nhưng những giải thuật này vẫn còn những hạn chế như cần dung lượng bộ nhớ lớn hoặc thời gian sửa lỗi khá nhiều. Trong bài báo này chúng tôi đề xuất giải thuật hiệu chỉnh lỗi, được đặt tên là RCorrector, dựa trên cấu trúc chỉ mục kmer nhằm phát hiện lỗi và sửa lỗi trực tiếp trên các read. So sánh với giải thuật SHREC trên 8 tập dữ liệu, RCorrector đạt được hiệu suất sửa lỗi thông qua hai đặc trưng specificity và sensitivity là tương đương với SHREC nhưng nhanh hơn SHREC từ 3 đến 7 lần.

1 GIỚI THIỆU

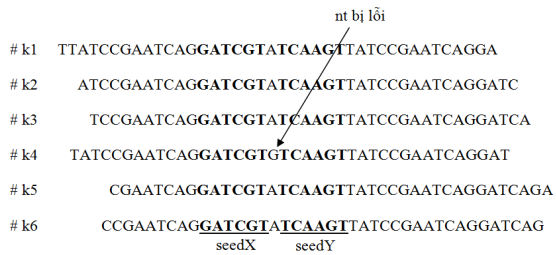
Xác lập trình tự ADN cho một gien của loài nào đó là một trong những công việc quan trọng trong lĩnh vực Sinh Tin học (Bioinformatics). Kỹ thuật xác lập trình tự thủ công đầu tiên được giới thiệu vào năm 1977 [6]. Với những cải tiến cho đến đầu những năm 1990, kỹ thuật xác lập trình tự có thể tạo ra các chuỗi ADN với độ dài từ 500bp đến 1000bp với độ chính xác 99%. Những năm gần đây kỹ thuật xác lập trình tự ADN đang trải qua một cuộc cách mạng. Đặc biệt là từ năm 2008 đến nay

với kỹ thuật xác lập trình tự song song lớn (massive), còn được gọi là NGS (Next Generation Sequencing) [4,7], cho phép thu được số lượng rất lớn dữ liệu trình tự ADN với tốc độ nhanh hơn và giá thành rẻ hơn so với các phương pháp xác lập trình tự trước đó. Chẳng hạn như công nghệ SOLEXA có thể tạo ra trên 10 Gbp trình tự ADN ngắn (short read 30-70 bp) trong khoảng thời gian 4 ngày với chi phí khoảng 60 đô la cho mỗi Mbp [7]. Với chi phí thấp và thời gian thực hiện nhanh nên ngày càng có nhiều dự án xác lập gien mới. Chẳng hạn phương pháp chuẩn đoán bệnh dựa theo dữ liệu cá nhân đã được đề nghị bởi Guy và *ctv* [1].

Sau khi xác lập được số lượng lớn các short read, gọi tắc là read, thì công việc tiếp theo là sắp dãy (assembly) các read này để thu được bộ gien hoàn chỉnh. Nhưng trong quá trình xác lập trình tự ADN của các kỹ thuật thế hệ mới này khả năng sinh lỗi cho từng nucleotide (nt) cao hơn. Tỷ lệ lỗi khoảng từ 1% đến 2% trên chiều dài của read [8]. Do số lượng các read sinh ra là rất lớn nên số lượng lỗi cũng rất nhiều. Các nucleotide lỗi phải được sửa lỗi để phục vụ cho việc sắp dãy lại thành một bộ gien hoàn chỉnh. Do đó cần phải có giải pháp sửa lỗi ở giai đoạn này nhằm chỉnh sửa các read bị lỗi. Các phương pháp sửa lỗi được đưa ra chủ yếu theo hai hướng sau đây. Hướng thứ nhất thực hiện sửa lỗi bằng cách xây dựng cây hậu tố (suffix tree) cho tất cả read. Phương pháp này được đề nghị bởi Schröder và *ctv* với tên gọi là giải thuật SHREC [8]. Để sửa lỗi thì trước hết SHREC sẽ xây dựng cây hậu tố cho tất cả các read. Tiếp theo là quá trình duyệt và phân tích các nút trên cây có hai nút con. Khi đó một nút con được xem là lỗi nếu trọng số (độ bao phủ) của nút đó nhỏ hơn trọng số mong đợi. Hướng thứ hai được đề nghị bởi Li và *ctv* với tên giải thuật là SOAP de novo [3]. Phương pháp này sử dụng đồ thị de Bruijn (de Bruijn graph) để sắp dãy các read thành các contig, các trình tự ADN có chiều dài lớn hơn các read. Các read được kiểm tra lỗi để sửa lỗi trước khi tiến hành sắp dãy.

Trong bài viết, chúng tôi đề xuất giải thuật sửa các nucleotide lỗi trong các trình tự ADN ngắn thu được từ kỹ thuật xác lập trình tự thế hệ mới. Giải thuật của chúng tôi sửa một nucleotide lỗi thành nucleotide đúng theo nguyên tắc biểu quyết theo số đông. Phương pháp sửa lỗi này hoạt động với sự hỗ trợ của cấu trúc chỉ mục kmer [10].

Phần tiếp theo của bài báo được tổ chức như sau: trong phần 2, chúng tôi sẽ trình bày ý tưởng chính của giải thuật sửa lỗi ở các chuỗi ADN ngắn do chúng tôi đề xuất. Quá trình thực nghiệm sẽ được trình bày trong phần 3 trước khi kết luận và hướng phát triển được trình bày trong phần 4.



Hình 1: Sửa lỗi dựa trên độ bao phủ của read

2 GIẢI THUẬT HIỆU CHỈNH LỖI K-MER

Trong phần này chúng tôi trình bày giải thuật sửa lỗi dựa trên kỹ thuật lập chỉ mục k-mer, được đặt tên là giải thuật RCorrector.

2.1 Nguyên tắc sửa lỗi

Giải thuật RCorrector thực hiện việc sửa lỗi một nucleotide nào đó của read dựa trên thông tin về độ bao phủ (coverage) khi xác lập các read từ một gien nào đó. Giả sử một gien có chiều dài là n và một tập R gồm m read $\{R_1,\ldots,R_m\}$ được xác lập từ một gien đó. Các read có chiều dài l. Độ bao phủ c được xác định qua công thức như sau: $c=(lm)/n$. Như vậy, nếu một gien có chiều dài n=10kbp và tập read R có 7000 read với độ dài của mỗi read l=50 thì độ bao phủ được tính như sau $c=(50x7000)/10000=35$. Giải thuật RCorrector chỉ có thể sửa lỗi tại một nucleotide nào đó của read khi độ bao phủ tại vị trí đó lớn hơn 2. Nguyên tắc sửa lỗi của giải thuật RCorrector dựa vào thông tin về độ bao phủ c của chuỗi con có độ dài bằng k, được gọi là k-mer [10] nằm trong c read như Hình 1. Trong Hình 1, chuỗi con có độ dài bằng 6 (6-mer) như là hai chuỗi seedX, seedY nằm trong 6 read tương ứng độ bao phủ $c=6$. Dựa vào sự giống nhau của các seedX và seedY trong 6 read này chúng tôi tiến hành kiểm tra các nucleotide (nt) nằm giữa hai seedY và seedY trong 6 read này. Nếu có sự khác biệt giữa các nt nằm giữa hai seed này thì chúng tôi tiến hành chỉnh sửa cho đúng. Chẳng hạn trong Hình 1 thì nucleotide của read thứ 4 (#k4) là G trong khi 5 nucleotide còn lại đều là A nên chúng tôi tiến hành sửa nt G thành A. Đây là nguyên tắc sửa lỗi biểu quyết theo số đông.

2.2 Giải thuật Rcorrector

Giải thuật RCorrector bao gồm ba giai đoạn chính: (1) lập chỉ mục cho tất cả các read của tập dữ liệu read vào; (2) sửa lỗi cho từng chỉ mục của INDEX theo nguyên tắc biểu quyết số đông; (3) loại bỏ các read không thể sửa lỗi.

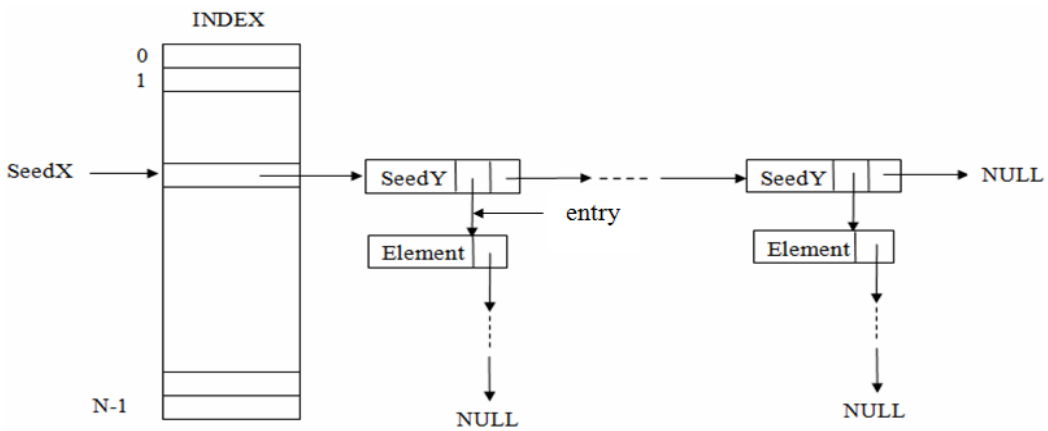
Giải thuật RCorrector

Đầu vào: tập tin chứa các read vào
Đầu ra: tập tin chứa các read ra
1: lập chỉ mục INDEX
2: **do** // {vòng lặp sửa lỗi}
3: **for each** mục của INDEX **do**
4: LR = danh sách read bị lỗi
5: **for each** phần tử read lỗi trong LR **do**
6: xóa các cặp chỉ mục trong INDEX
7: sửa lỗi nucleotide của read
8: chèn các cặp seed mới vào INDEX
9: **end for**
10: **end for**
11: **while** hết lỗi hoặc số lỗi không đổi
12: loại bỏ các read không thể sửa lỗi
13: xuất ra tập tin kết quả

2.2.1 *Lập chỉ mục INDEX*

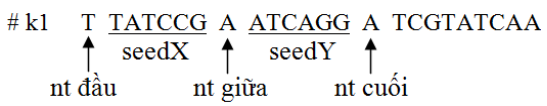
Để có thể xác định được các seedX và seedY giống nhau với độ bao phủ *c* chúng tôi sử dụng cấu trúc chỉ mục INDEX như Hình 2.

Cấu trúc chỉ mục này là một mảng được cấp phát động gồm *N* phần tử, với *N* được xác định dựa vào chiều dài của seedX. Gọi *s* là chiều dài của seedX thì số phần tử của cấu trúc INDEX là $N = 4^s$ bởi vì mỗi nucleotide của seedX ðược mã hóa bằng 2 bít nhị phân như sau *A=00, C=01, G=10* và *T=11*. Mỗi phần tử của mảng là một con trỏ, trỏ đến các node chứa seedY. Mỗi node của seedY có hai con trỏ: một tương ứng danh sách theo chiều ngang (Hnext) trỏ đến các seedY khác nhau, con trỏ còn lại tương ứng với danh sách theo chiều dọc (Vnext) chứa danh sách các phần tử có cùng giá trị seedY. Mỗi danh sách tương ứng với một seedY theo chiều dọc được gọi là một entry. Mỗi phần tử element của entry lưu thông tin gồm năm thành phần như sau: {#, pos, c_1, c_2, c_3}. Trong đó #: số thứ tự của read trong tập dữ liệu; pos: vị trí của nucleotide giữa seedX và seedY của read; c1, c2, c3 tương ứng lần lượt là ba nt kề đầu seedX, nt kề giữa seedX và seedY và nt kề cuối của seedY như Hình 3.



Hình 2: Cấu trúc chỉ mục INDEX

Quá trình lưu bộ element gồm 5 thành phần của một cặp {seedX, seedY} của read trong tập dữ liệu đầu vào được thực hiện như sau: đầu tiên cần xác định vị trí của seedX trong cấu trúc INDEX. Tiếp theo từ vị trí này RCorrector tiến hành xác định vị trí theo chiều ngang (Hnext) để kiểm tra sự tồn tại của chuỗi seedY này hay chưa. Kết quả tìm kiếm theo chiều ngang sẽ có hai khả năng:



Hình 3: Cấu trúc các thành phần của 1 element.

– Nếu seedY đã tồn tại, bộ element sẽ được chèn vào danh sách liên kết tương ứng theo chiều dọc (entry) của seedY này. Danh sách được trỏ bởi Vnext và được sắp xếp theo thứ tự tăng dần của chỉ số read nhằm giúp cho quá trình tìm kiếm khi xóa và lập chỉ mục mới hiệu quả hơn.

– Nếu không tìm thấy, thực hiện chèn thêm một nút mới ở đầu danh sách theo chiều ngang và gán giá trị seedY mới. Đồng thời chèn phần tử element vào nút mới này bởi con trỏ theo chiều dọc (Vnext).

Như vậy với một read có chiều dài *l* và kích thước của seed (X và Y) là *s* thì sẽ có tổng cộng $l-(2\times s+2)$ bộ element được đưa vào cấu trúc chỉ mục INDEX.

2.2.2 Sửa lỗi ở các read

Sau khi lập chỉ mục INDEX toàn bộ các read trong tập dữ liệu đầu vào, RCorrector tiến hành duyệt qua từng mục trong cấu trúc INDEX để tiến hành sửa lỗi. Tại một chỉ mục nào đó tương ứng với seedX, dữ liệu trong mục này là danh sách ngành với nhiều entry. Mỗi entry tương ứng với một seedY được biểu diễn bởi con trỏ Hnext. Dữ liệu trong mỗi entry được minh họa như Bảng 1. Cột đầu tiên là số thứ tự (#) của read được trong tập dữ liệu vào và được xếp theo thứ tự tăng dần; cột thứ hai là vị trí (pos) của nt nằm giữa seedX và seedY; bộ ba kí tự ở cột thứ ba tương ứng là ba nt kề đầu seedX, kề giữa seedX và seedY, kề cuối của seedY.

Bảng 1: Lỗi trong entry có thể được được sửa

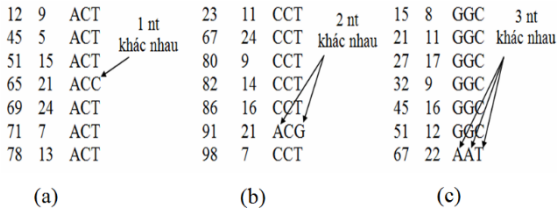
Số thứ tự (#k)	Vị trí nt giữa seedX và seedY (pos)	Ba kí tự kề trước, giữa và cuối của {seedX, seedY}
12	9	ACT
45	5	ACT
51	15	ACC ← nt bị lỗi
65	21	ACT
69	24	ACT
71	7	ACT
78	13	ACT

Cột bộ ba ký tự kề trước, kề giữa, kề cuối của seedX và seedY trong mỗi entry chính là cơ sở để giải thuật RCorrector tiến hành sửa lỗi theo nguyên tắc biểu quyết theo số đông. Chẳng hạn trong Bảng 1, cột này chỉ ra rằng read có số thứ tự 51 ở hàng thứ 3 có nt kề cuối của seedY là "C" được xem là bị lỗi do khác biệt với "T" của các bộ ba còn lại. Do đó, nt "C" sẽ được sửa lại thành "T" theo nt đúng của entry. Trong entry này các bộ ba ký tự đầu, giữa và cuối đồng nhất với nhau và chỉ có một ký tự "C" của dòng thứ ba là khác biệt nên dễ dàng được sửa lại thành "T".

Bảng 2: Lỗi trong entry không thể sửa được

Số thứ tự (#k)	Vị trí nt giữa seedX và seedY (pos)	Ba kí tự kề trước, giữa và cuối của {seedX, seedY}
23	11	GGT
67	24	GGT
80	9	GGT
82	14	GCT ← không thể sửa lỗi
86	16	CCT
91	21	CCT
98	7	CCT

Nhưng các bộ ba ký tự của một entry khác như ở Bảng 2 thì có 2 nhóm đại diện (GGT và CCT). Trong khi đó read có số thứ tự là 82 ở hàng thứ 4 có bộ 3 nt là "GCT". Nếu so sánh "GCT" với nhóm thứ nhất là "GGT" thì sẽ sửa nt kề giữa "C" thành "G". Còn nếu so sánh "GCT" với nhóm thứ hai là "CCT" thì sẽ sửa nt kề trước "G" thành "C". Vì vậy, RCorrector không thể xác định được read (số thứ tự 82) thuộc nhóm nào. Trong trường hợp này giải thuật sẽ không thực hiện sửa lỗi bởi vì xác suất sửa lỗi sai sẽ rất cao. Do đó, RCorrector phải kiểm tra sự đồng nhất của các bộ ba nt, được gọi là 3 nt đúng, trong entry với yêu cầu chỉ có một bộ 3 ký tự có sự khác biệt với các bộ ba ký tự còn lại như ở Hình 4. Hình 4 này minh họa ba khả năng xảy ra: (a) chỉ có một nt sai so với 3 nt đúng của entry; (b) hai nt sai so với 3 nt đúng của entry; (c) cả ba nt đều sai so với 3 nt đúng của entry. RCorrector chỉ tiến hành sửa lỗi cho trường hợp đầu tiên.



Hình 4: Các trường hợp khác biệt với bộ nt đúng

Khi tiến hành sửa lỗi cho một element của entry thì lỗi này có thể là xảy ra ở nt kề trước seedX, hoặc nt kề giữa seedX và seedY hoặc nt kề cuối với seedY. Trước khi tiến hành sửa lỗi một nt, RCorrector tiến hành quá trình xóa các bộ element của read được sửa lỗi và cập nhập lại các bộ element cho read này phụ thuộc vào vị trí của nt được sửa lỗi như sau:

– Nếu lỗi tại ký tự kề trước của seedX thì quá trình xóa các bộ element và lập các bộ element mới chỉ thực hiện cho các cặp {seedX, seedY} liên quan đến nt bị lỗi. Tổng số bộ element phải sửa phụ thuộc vào vị trí của nt lỗi trong read. Gọi e là vị trí của nt lỗi, s là kích thước của seed. Nếu $e \geq (2s+3)$ tổng số bộ element phải sửa là $(2s+3)$ ngược lại thì số bộ element phải sửa là $e+1$.

– Nếu sửa lỗi tại ký tự kề giữa seedX và seedY thì quá trình xóa các bộ element và lập các bộ element mới cho các cặp seedX và seedY phía trước và phía sau có liên quan đến nt bị lỗi. Tùy theo vị trí xảy ra lỗi trong read mà số bộ element có liên quan từ $s+3$ đến $2s+3$.

– Nếu sửa lỗi tại ký tự kề cuối của seedY thì số bộ element phải chỉnh sửa sẽ là một nếu $e=l-1$, trong đó l là chiều dài của read, ngược lại thì số bộ element được cập nhật nhiều nhất là $2s+3$.

2.2.3 Loại bỏ các read không thể sửa lỗi

Sau mỗi vòng lặp, số lỗi được phát hiện sẽ giảm dần và có thể sau một số vòng lặp thì giải thuật RCorrector sẽ không phát hiện thêm lỗi nào nữa. Khi đó giải thuật RCorrector sẽ kết thúc và không có read nào được loại bỏ khỏi tập dữ liệu đầu ra. Nhưng cũng có thể xảy ra trường hợp sau một số vòng lặp số lỗi mà giải thuật RCorrector phát hiện bị lặp lại. Điều này là vì một số nt của read có thể bị sửa nhầm thành nt sai. Do đó, sau khi cập nhập lại với nt mới thì các cặp {seedX, seedY} vừa được đưa vào bị trùng vào vị trí khác trên gien nhưng vị trí này không phải là vị trí đúng của read. Ở vòng lặp tiếp theo, nt được sửa theo entry này nhưng trong vòng lặp sau nt đó được sửa lại thành nt khác theo entry khác. Cho nên quá trình này bị lặp đi lặp lại gây ra số lỗi do giải thuật phát hiện cũng sẽ bị lặp lại.

Khi số lỗi do giải thuật phát hiện sau một số vòng bị lặp lại thì tất cả các read liên quan đến các lỗi này đều không thể được sửa lỗi. Số read này sẽ được loại bỏ khỏi tập dữ đầu ra.

3 KẾT QUẢ THỰC NGHIỆM

3.1 Các tiêu chí đánh giá giải thuật

Để đánh giá giải thuật RCorrector chúng tôi chọn các tiêu chí để đánh giá về thời gian thực thi của giải thuật và hai đặc trưng thể hiện khả năng sửa lỗi và độ tin cây là: Sensitivity và Specificity. Hai đặc trưng này được tính như sau:

Sensitivity = TP / (TP+FN)

Specificity = TN / (TN+FP)

Trong đó:

– True Positive (TP): read có nt lỗi và nt này được sửa lại thành đúng bởi giải thuật.

– False Positive (FP): read có nt đúng nhưng được sửa lại thành nt sai bởi giải thuật.

– True Negative (TN): read có nt đúng và vẫn được giữ nguyên nt đúng.

– False Negative (FN): read có nt lỗi và vẫn còn giữ nguyên nt lỗi sau khi thực hiện giải thuật.

Đặc trưng Sensitivity thể hiện khả năng sửa lỗi của giải thuật. Giải thuật sửa lỗi càng hiệu quả (Sensitivity càng gần bằng 100%) khi số read vẫn còn bị lỗi (FN) nhỏ hơn nhiều so với số read bị lỗi mà đã được sửa đúng (TP). Trong khi đó, đặc trưng Specificity thể hiện khả năng này thể hiện độ tin cậy của giải thuật sửa lỗi.

3.2 Các tập dữ liệu

Giải thuật RCorrector được đánh giá dựa trên các tập dữ liệu giả lập (simulated datasets). Các tập dữ liệu giả lập được xây dựng từ 4 tập dữ liệu được tải từ trang web NCBI: Saccharomyces cerevisiae chromosome 5 (S.cer5), chromosome 7 (S.cer7), gien vi khuẩn Haemophilus influenzae (H.inf) và Staphylococcus Aureus (S.aureus). Kích thước của 4 tập dữ liệu này từ 0.58 Mbp tới 2.8 Mbp. Với mỗi tập dữ liệu đã chọn, chúng tôi tạo ra hai tập dữ liệu read giả lập với tỷ lệ lỗi tương ứng là 1% và 2%. Thông tin chi tiết các tập dữ liệu read được trình bày ở Bảng 3.

Bảng 3: Các tập dữ liệu dùng để chạy thực nghiệm

ID	Reference genome (Genbank ID)	Genome length	Error - rate	Coverage	Read length	Number of reads
A1	S.cer 5 (NC_001137.3)	0.58M	1%	35	35	0.58M
A2			2%			
B1	S.cer 7 (NC_001139.9)	1.1M	1%			1.1M
B2			2%			
C1	H.inf (NC_007146.2)	1.9M	1%			1.9M
C2			2%			
D1	S.aureus (NC_003923.1)	2.8M	1%			2.8M
D2			2%			

3.3 Kết quả thực nghiệm

Để đánh giả giải thuật RCorrector, chúng tôi tiến hành chạy thực nghiệm giải thuật RCorrector và giải thuật SHREC với 8 tập dữ liệu ở Bảng 3 trên máy PC sử dụng CPU intel Genuine 2.13GHz với bộ nhớ RAM 3GB chạy trên hệ điều hành Linux Ubuntu 10. Giải thuật RCorrector được cài đặt bằng ngôn ngữ lập trình C. Quá trình đánh giá gồm có hai phần: phần thứ nhất tiến hành đánh giá giải thuật RCorrector với nhiều kích thước seed khác nhau. Từ đó chọn kích thước của seed tối ưu nhất để làm giá trị mặc định cho giải thuật RCorrector. Phần thứ hai tiến hành đánh giá so sánh giữa giải thuật RCorrector và giải thuật SHREC về thời gian chạy cũng như hai đặc trưng sửa lỗi Sensitivity và Specificity.

3.3.1 Kết quả đánh giá giải thuật Rcorrector với các chiều dài seed khác nhau

Để kiểm tra khả năng sửa lỗi của giải thuật RCorrector, chúng tôi tiến hành thực nghiệm nhằm đánh giá giải thuật với nhiều kích thước của seed khác nhau (từ 6 nt đến 10 nt) trên tập dữ liệu C1. Với chiều dài seed nhỏ hơn 6 thì khả năng xảy ra lặp lại (repeat) của các trình tự của seedX và seedY tại nhiều vị trí trên gien sẽ làm cho giải thuật sửa lỗi kém hiệu quả về mặt thời gian vì kích thước lưu trữ trong từng chỉ mục của INDEX rất lớn. Ngược lại, đối với chiều dài seed lớn hơn 10 kích thước lưu trữ trong từng chỉ mục INDEX rất ít nên khả năng phát hiện lỗi sẽ rất thấp. Vì vậy, chiều dài seed nhỏ hơn 6 hoặc lớn hơn 10 không được chúng tôi dùng trong thực nghiệm để đánh giá. Kết quả của giải thuật RCorrector với độ dài seed từ 6 đến 10 được thể hiện ở Bảng 4.

Bảng 4: Thời gian chạy (tính bằng giây) và hai đặc trưng sửa lỗi

Chiều dài seed	6	7	8	9	10
Thời gian	7385	2331	676	267	158
Sensitivity	76.23%	95.02%	95.39%	92.24%	87.66%
Specificity	99.67%	99.89%	99.93%	99.95%	99.96%

Bảng 4 chỉ ra rằng khi tăng chiều dài seed thì thời gian thực hiện của giải thuật giảm dần. Điều này được giải thích như sau: đối với một read, khi tăng chiều dài seed số lượng cặp {seedX, seedY} được tạo ra ít hơn cho nên thời gian xử lý lỗi sẽ nhanh hơn. Đặc trưng sửa lỗi Specificity của giải thuật RCorrector thể hiện giá trị Specificity cao (đạt trên 99.6%) đối với tất cả các seed. Đặc trưng Sensitivity đạt giá trị tốt nhất với chiều dài seed bằng 8. Khi chiều dài seed tăng dần từ 8 đến 10 thì chỉ số Sensitivity giảm xuống. Điều này là do khi chiều dài seed càng lớn thì số cặp {seedX, seedY} trong chỉ mục của INDEX càng ít. Do đó khả năng phát hiện và sửa lỗi của giải thuật Rcorrector sẽ giảm.

Tóm lại, với chiều dài seed bằng 8 là lựa chọn tối ưu cho giải thuật RCorrector bởi vì thời gian thực sửa khá nhanh (676s) nhưng đặc trưng sửa lỗi Sensitivity đạt giá trị tốt nhất (95.39%). Vì vậy, kích thước của seed bằng 8 được chọn làm giá trị mặc định cho giải thuật RCorrector.

3.3.2 So sánh giải thuật RCorrector với giải thuật SHREC

Tiếp theo chúng tôi trình bày kết quả đánh giá giải thuật RCorrector bằng cách so sánh với giải thuật SHREC về thời gian thực hiện và hai đặc trưng Sensitivity, Specificity. Cả hai giải thuật SHREC và RCorrector được tiến hành chạy thực nghiệm trên 8 tập dữ liệu với tham số mặc định.

a. Về thời gian chạy chương trình

Thời gian thực hiện của hai giải thuật được thể hiện ở Bảng 5. So với SHREC thì RCorrector chạy nhanh hơn từ 3 từ 7 lần.

Xét trên một tập dữ liệu, ví dụ S.cer 5, thời gian thực hiện của RCorrector phụ thuộc vào tỷ lệ lỗi ở các read. Nếu tỷ lệ lỗi càng cao thì RCorrector cần nhiều thời gian chỉnh sửa lỗi. Khi tỷ lệ lỗi tăng thì tổng số lỗi mà giải thuật phát hiện cũng sẽ tăng lên, do đó cần nhiều thời gian hơn để sửa lỗi. Thao tác sửa lỗi của SHREC là duyệt cây và trong lúc duyệt nếu phát hiện lỗi thì tiến hành sửa lỗi. Sau khi sửa lỗi SHREC phải cập nhật lại cây.

Bảng 5: Thời gian chạy (giây) của SHREC và Rcorrector trên 8 tập dữ liệu

Tập dữ liệu	A1	A2	B1	B2	C1	C2	D1	D2
SHREC	354	463	1046	1102	4734	4948	8544	10800
RCorrector	75	112	223	364	676	1221	1771	3423
Speedup	5	4	5	3	7	4	5	3

b. Về hiệu suất hiệu chỉnh sửa lỗi

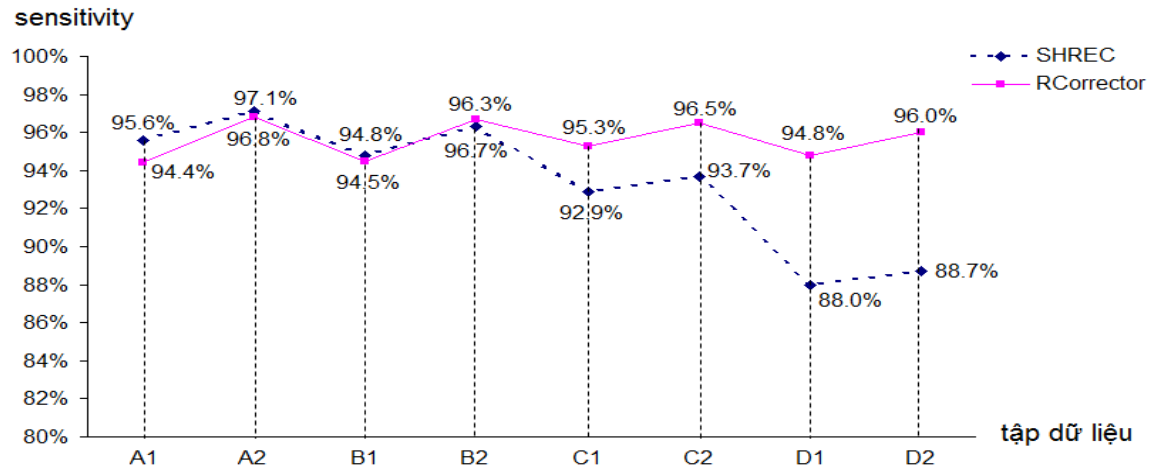
Hai đặc trưng sửa lỗi Sensitivity và Specificity của giải thuật RCorrector và SHREC được tính thông qua các đại lượng True Positive (TP), False Positive (FP), False Negative (FN) và True Negative (TN) được thể hiện ở Bảng 6 và Bảng 7. Trước hết, xét đặc trưng Specificity của hai giải thuật (cột cuối cùng của Bảng 6&7). Đặc trưng này thể hiện độ tin cậy của một giải thuật sửa lỗi. Nó phụ thuộc vào số read bị sửa nhầm (FP) so với số read đúng (TN). Đối với giải thuật SHREC, đặc trưng Specificity đạt tốt nhất (100%) trên tất cả các tập dữ liệu. Điều này là do số read bị sửa nhầm của giải thuật rất nhỏ. Đối với RCorrector, số read bị sửa nhầm có nhiều hơn so với SHREC nhưng cũng không đáng kể trên tổng số read đúng. Do đó, Specificity đạt cũng rất cao (hơn 99.6%) đối với tất cả các tập dữ liệu. Nói chung, cả hai giải thuật có độ tin cậy rất cao. Đặc điểm này là rất cần thiết đối với một giải thuật sửa lỗi.

Bảng 6: Các đặc trưng Sensitivity và Specificity trên 8 tập dữ liệu của Rcorrector

	TP	FP	FN	TN	Sensitivity	Specificity
A1	125482	312	7344	443988	0.944	0.999
A2	279094	201	9179	288522	0.968	0.999
B1	237351	1818	13636	839702	0.945	0.997
B2	527768	1091	17942	544951	0.967	0.998
C1	418988	1006	20265	1474961	0.953	0.999
C2	922982	495	32757	958333	0.965	0.999
D1	614505	3867	33172	2172240	0.948	0.998
D2	1354444	2516	55459	1409631	0.960	0.998

Bảng 7: Các đặc trưng Sensitivity và Specificity trên 8 tập dữ liệu của SHREC

	TP	FP	FN	TN	Sensitivity	Specificity
A1	127034	0	5803	444000	0.956	1
A2	280166	0	8143	288528	0.971	1
B1	238058	2	13015	839829	0.948	1
B2	525921	1	19945	545036	0.963	1
C1	408568	0	30868	1475016	0.929	1
C2	895823	0	60275	958354	0.937	1
D1	570705	1	77419	2172300	0.880	1
D2	1252574	1	158178	1409672	0.887	1



Hình 5: Đặc trưng Sensitivity của SHREC và RCorrector trên các tập dữ liệu

Đặc trưng Sensitivity của hai giải thuật được thể hiện ở cột thứ 6 của Bảng 6&7 và Hình 5. Đây là đặc trưng thể hiện khả năng phát hiện lỗi của các giải thuật. Với hai tập dữ liệu nhỏ (A1 và A2) thì Sensitivity của giải thuật SHREC tốt hơn giải thuật RCorrector. Như đối với các tập dữ liệu còn lại thì Sensitivity của RCorrector tốt hơn giải thuật SHREC. Điều này được giải thích như sau: Với dữ liệu nhỏ thì số lượng seed trong các mục trong cấu trúc INDEX không nhiều nên có thể không đạt ngưỡng để tiến hành sửa lỗi theo nguyên tắc biểu quyết theo số đông, còn đối với các tập dữ liệu lớn hơn thì nguyên tắc biểu quyết theo số đông hoạt động tốt hơn quá trình sửa lỗi của Rcorrector tốt hơn.

4 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi vừa trình bày trong bài báo giải thuật RCorrector sửa lỗi ở các chuỗi ADN ngắn thu được từ kỹ thuật xác lập trình tự thế hệ mới. Giải thuật RCorrector quyết định sửa một nucleotide lỗi thành nucleotide đúng theo nguyên tắc biểu quyết theo số đông. Phương pháp sửa lỗi này của RCorrector hoạt động hiệu quả với sự hỗ trợ của cấu trúc chỉ mục INDEX dựa trên kỹ thuật kmer. Kết quả thực nghiệm trên 8 tập dữ liệu giả lập cho thấy rằng giải thuật RCorrector nhanh hơn giải thuật SHREC từ 3

đến 7 với hai đặc trưng sửa lỗi Sensitivity và Specificty của hai giải thuật RCorrector và SHREC tương đương nhau. Trong tương lai, chúng tôi nghiên cứu để song song hóa giải thuật RCorrector trên PC cluster nhằm giảm thời gian sửa lỗi cho các tập dữ liệu lớn hơn.

TÀI LIỆU THAM KHẢO

1. Guy Haskin Fernald, Emidio Capriotti, Roxana Daneshjou,Konrad J. Karczewski, and Russ B. Altman. Bioinformatics challenges for personalized medicine. Bioinformatics (2011) 27 (13): 1741-1748
2. Li R., Zhu H., Ruan J., *et al*., De novo assembly of human genomes with massively parallel short read sequencing, Genome Research, volume 20, number 2, pp: 265–272, 2010.
3. Monya Baker, De novo genome assembly: what every biologist should know, Nature, volume 9, pp:333–337, 2012.
4. Pop M., Salzberg S. L., Bioinformatics challenges of new sequencing technology, Trends in Genetics, 24 (3), pp:142-149, 2008.
5. Salmela L., Correction of sequencing errors in a maxed set of reads. Bioinformatics 26(10) pp:1284-1290, 2010.
6. Sanger, F. *et al*. Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265, 687–695 (1977).
7. Shendure J. and Ji H., Next-generation DNA sequencing, Nature biotechnology, volume 26, number 10, 1135-1145, 2008.
8. Schröder J., Schröder H., Simon J. P., Sinja R. and Schmidt B., SHREC: A short-read error correction method, Genome Analysis, volume 25, number 17, 2157-2163, 2009.
9. Tammi M. T., Arner E., Kindlund E., Andersson B., Correcting errors for shotgun sequencing, Nucleic Acid Research, 31, pp:4663-4672, 2003.
10. V.H. Nguyen, D. Lavenier, PLAST: parallel local alignment search tool for database comparison, BMC Bioinformatics 2009 10(329).

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

THIẾT KẾ HỆ THỐNG THANH TOÁN CÁC DỊCH VỤ CÔNG CỘNG ỨNG DỤNG CÔNG NGHỆ RFID VÀ NFC

Đặng Vũ Minh Dũng[1], Đoàn Quốc Nam[2] và Lương Vinh Quốc Danh[3]

[1] *Ngành Kỹ Thuật Máy tính, Khoa Công Nghệ, Trường Đại học Cần Thơ*
[2] *Ngành Kỹ Thuật Điều khiển, Khoa Công Nghệ, Trường Đại học Cần Thơ*
[3] *Bộ môn Điện tử Viễn thông, Khoa Công Nghệ, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Design of an e-payment system for public services using RFID and NFC technologies

Từ khóa:
Thanh toán điện tử, dịch vụ công cộng, RFID, NFC

Keywords:
E-payment, public services, RFID, NFC

ABSTRACT

Radio-Frequency Identification (RFID) and Near Field Communication (NFC) technologies applied in e-payment systems have been employed in many countries around the world for a variety of areas such as paying fees in public transportations, buying foods at fast food restaurants, at super markets, and buying soft drink at vending machines. In recent years, these technology solutions have been used in some regions in Vietnam through paying bills at parking services, vending machines, school and company canteens… These technologies make our life easier and more convenient for consumers, save people's times, reduce cash transactions, centrally manage public services, and modernize our daily life.

This paper presents the prospects of applying the RFID and NFC technologies to the e-payment methods illustrated through the design of an e-payment system for motorcycle parking services. The results of this work have demonstrated the feasibility of implementing the system's hardware in Vietnam.

TÓM TẮT

Công nghệ RFID (Radio-Frequency Identification) và NFC (Near Field Commnication) ứng dụng trong thanh toán điện tử là một giải pháp đã được đưa vào sử dụng tại nhiều nước trên thế giới trong các lĩnh vực như thu phí giao thông công cộng, thanh toán tiền tại các hiệu ăn nhanh, siêu thị, máy bán hàng tự động. Trong vài năm gần đây, giải pháp này cũng đã và đang được áp dụng tại một số địa phương ở Việt Nam qua các dịch vụ thanh toán phí giữ xe, máy bán hàng tự động, dịch vụ căng-tin trường học, công sở... Giải pháp này mang đến sự tiện lợi và nhanh chóng trong thanh toán, giúp tiết kiệm thời gian, giảm lượng tiền mặt lưu thông trên thị trường, quản lý các dịch vụ tập trung đồng thời tạo nên bộ mặt hiện đại cho xã hội.

Mục tiêu của bài viết nhằm giới thiệu khả năng ứng dụng công nghệ RFID và NFC để giải quyết bài toán thanh toán tại nơi công cộng thông qua một ứng dụng cụ thể là thanh toán chi phí giữ xe. Kết quả nghiên cứu cũng cho thấy việc thiết kế, chế tạo các mạch điện của hệ thống là hoàn toàn khả thi với kỹ thuật hiện tại ở Việt Nam.

1 GIỚI THIỆU VỀ RFID VÀ NFC

1.1 Đặc điểm kỹ thuật

NFC là công nghệ kết nối không dây tầm ngắn sử dụng cảm ứng điện từ trường để thực hiện kết nối giữa các thiết bị khi có sự tiếp xúc trực tiếp hay để gần nhau. NFC hoạt động trên dải băng tần 13,56 MHz và có tốc độ truyền tải dữ liệu tối đa 424 Kbps [1]. Do khoảng cách truyền dữ liệu khá ngắn nên giao dịch bằng công nghệ NFC được xem là an toàn.

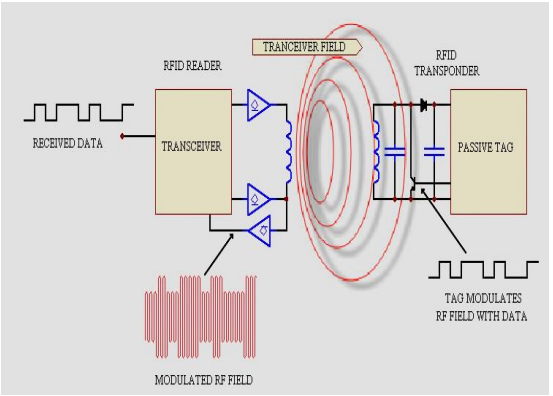
Hình 1: Ứng dụng thẻ NFC FeliCa của Sony [2]

Các ứng dụng phổ biến của thẻ NFC được mô tả trong Hình 1 thuộc nhiều lĩnh vực khác nhau như vé điện tử, vé xe, thẻ chứng minh thư, thẻ đa chức năng...

RFID là phương thức nhận dạng tự động, cho phép lưu trữ và lấy dữ liệu từ xa dựa vào cộng hưởng sóng vô tuyến. Thẻ RFID được gắn kèm vào sản phẩm. Hệ thống RFID gồm hai phần: thẻ tag có gắn chip silicon cùng ăng-ten và phần thứ 2 là bộ đọc giao tiếp với thẻ tag và truyền dữ liệu đến hệ thống máy tính trung tâm. Tần số của RFID đa dạng hơn thẻ NFC, các dải tần của RFID là 125 KHz, 13.56 MHz, dải tần UHF, và dải tần microwave. Hình 2 mô tả hoạt động của thẻ RFID giữa thẻ Tag và đầu đọc dựa trên nguyên lý cảm ứng giữa điện từ [3].

Đối với RFID sử dụng cho việc thanh toán điện tử, dải tần số HF 13.56 MHz thường được sử dụng. Khoảng cách giữa đầu đọc và thẻ tag là dưới 10 cm nên an toàn về mặt bảo mật thông tin. Ngoài ra, dải tần 13.56 MHz có thể sử dụng chung một đầu đọc cho cả RFID và NFC.

Ngày nay, công nghệ NFC đã được tích hợp trên điện thoại thông minh (smartphone) nên trong tương lai, công nghệ này có triển vọng phát triển rất lớn.



Hình 2: RFID hoạt động trên nguyên lý cảm ứng điện từ [4]

1.2 Việc ứng dụng RFID và NFC ở trong và ngoài nước

– **Nước ngoài:** Trên thế giới, việc áp dụng công nghệ RFID và NFC vào việc thanh toán các dịch vụ công cộng đã được thực hiện ở nhiều quốc gia khác nhau và đã đem lại nhiều thành công. Lợi ích của công nghệ này mang lại là sự tiện lợi cho người dùng, tiết kiệm được thời gian và chi phí nhân công, đồng thời xây dựng một bộ mặt hiện đại cho xã hội. Tại Singapore, công nghệ RFID EZ-LINK được sử dụng trong việc thanh toán phí giao thông công cộng [5]. Tại Hồng Kông, dịch vụ thanh toán đa năng Octopus được sử dụng trong việc thanh toán phí giao thông công cộng, chi phí dịch vụ mua sắm, ăn uống [6]. Đặc biệt, Nhật Bản là nước đầu tiên trên thế giới áp dụng thành công công nghệ NFC vào việc thanh toán ở quy mô lớn với dịch vụ FeliCa Network của nhà mạng NTT Docomo dùng những chiếc điện thoại có tích hợp công nghệ NFC [7].

– **Trong nước:** Ở Việt Nam với dân số xấp xỉ 90 triệu người ước tính vào năm 2013 [8], việc sử dụng các dịch vụ công cộng như dịch vụ giữ xe, thức ăn nhanh, dịch vụ vận chuyển và giao thông có nhu cầu rất lớn. Đây chính là điều kiện rất thuận lợi để áp dụng giải pháp thanh toán điện tử. Hiện tại, công nghệ RFID và NFC cũng bắt đầu được áp dụng trong việc thanh toán chi phí giữ xe tại các siêu thị, hệ thống đóng/mở cổng tự động tại các bãi đỗ xe ô tô, kiểm soát hiện diện của nhân việc trong công ty… ở một số tỉnh, thành phố.

1.3 Triển vọng ứng dụng công nghệ RFID và NFC tại Việt Nam

Những vấn đề phát sinh khi thanh toán bằng tiền mặt như giá trị thanh toán nhỏ và khách hàng sử dụng tiền mặt có mệnh giá lớn, trong khi người bán hàng không có đủ tiền lẻ để trả lại cho khách

thì số tiền cần phải trả thường được làm tròn. Ngoài ra, việc hoàn trả tiền lẻ cho khách hàng làm cho quá trình thanh toán kéo dài dẫn đến việc mất thời gian. Riêng đối với dịch vụ giữ xe, việc này có thể dẫn đến tình trạng ùn tắc xe vào giờ cao điểm. Vì vậy, việc áp dụng giải pháp thanh toán điện tử với công nghệ RFID hoặc NFC là một giải pháp có thể khắc phục được các vấn đề trên.

Giải pháp thanh toán điện tử qua công nghệ RFID và NFC có thể áp dụng ở các khu vực từ trường học đến công sở, rộng hơn là thanh toán phí xe buýt công cộng. Ngoài ra công nghệ này có thể được áp dụng làm thẻ khám chữa bệnh và thanh toán chi phí khám chữa bệnh tại các bệnh viện. Trường học và công sở là những nơi thuận lợi nhất để áp dụng các giải pháp thanh toán điện tử này do những nơi này tập trung nhiều dịch vụ công cộng như dịch vụ giữ xe, dịch vụ ăn uống tại các căng-tin, máy bán hàng tự động, in ấn, dịch vụ photocopy. Một thuận lợi nữa là những nơi này tập trung nhiều người trẻ, năng động vì thế việc tiếp cận giải pháp sẽ dễ dàng và nhanh chóng hơn.

2 HỆ THỐNG THANH TOÁN ĐIỆN TỬ SỬ DỤNG CÔNG NGHỆ RFID VÀ NFC

2.1 Phương pháp nghiên cứu

2.1.1 Cấu tạo hệ thống thanh toán điện tử

Trong nghiên cứu này, hệ thống thanh toán điện tử sử dụng công nghệ RFID và NFC có cấu tạo gồm 3 phần chính [9]: Server, client và thẻ RFID hoặc NFC.

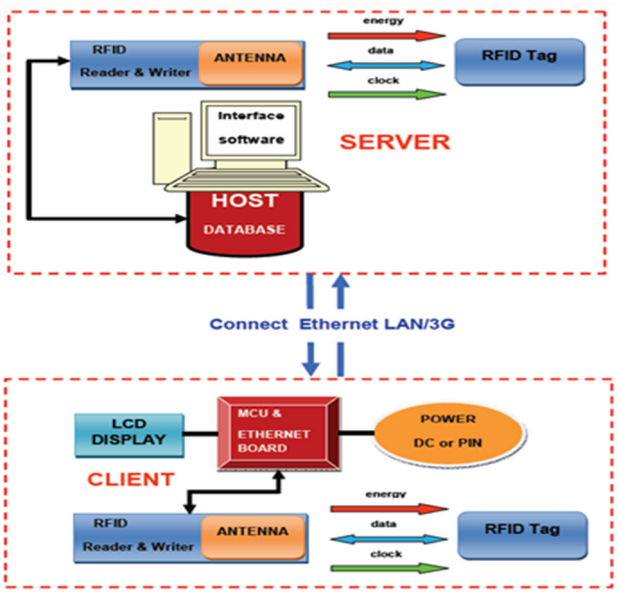
– **Server:** Hệ thống máy chủ có nhiệm vụ lưu trữ cơ sở dữ liệu, thông tin người dùng, số tiền trong tài khoản, lịch sử giao dịch; Nhận và xử lý dữ liệu từ các client và gửi trở lại kết quả giao dịch cho client.

– **Client:** Là nơi đọc thông tin của thẻ thanh toán RFID hoặc NFC và gửi dữ liệu về máy chủ để xử lý. Thông tin giao dịch của chủ thẻ được server gửi trả về client sẽ được hiển thị trên màn hình LCD tại client.

– **Thẻ thanh toán:** Mỗi người dùng sẽ được cấp một thẻ RFID hoặc NFC, thẻ này tương ứng với một tài khoản được cấp trên hệ thống server. Người sử dụng muốn thực hiện thanh toán phải dùng thẻ được cấp quét qua đầu đọc ở client.

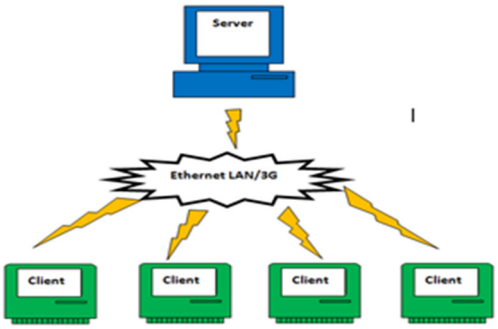
Hình 3 trình bày sơ đồ khối liên kết của một hệ thống bao gồm một máy chủ server và một máy khách client. Phía server sẽ bao gồm một máy chủ lưu trữ cơ sở dữ liệu database, được trang bị một đầu đọc thẻ RFID và NFC để đọc ID thẻ. Một phần mềm được tạo ra trên máy chủ để xử lý các công việc bao gồm đọc ID thẻ và lưu vào cơ sở dữ liệu, sửa chữa thông tin người dùng, nạp tiền vào tài khoản, lưu trữ lịch sử thanh toán, kiểm tra hệ thống, kết nối qua Ethernet LAN/3G với các client.

Phía các client là một module bao gồm vi điều khiển MCU, một module enthernet, một đầu đọc RFID và NFC, một màn hình LCD, một keyboard. Khi người dùng quét thẻ qua đầu đọc, các ID người dùng sẽ được client gửi đến server qua kết nối ethernet. Các client chờ nhận kết quả xử lý từ server và hiển thị thông tin giao dịch của lần thanh toán đó ra màn hình LCD cho người dùng biết.



Hình 3: Hệ thống thanh toán qua công nghệ RFID và NFC

Để tạo ra một hệ thống thanh toán ở khu vực rộng, hệ thống sẽ gồm có một server chính và các client tương ứng với các điểm thanh toán, các client kết nối với server qua Ethernet LAN/3G để trao đổi dữ liệu.



Hình 4: Hệ thống thanh toán nhiều client

Mô hình tổng quát của hệ thống sẽ bao gồm nhiều client kết nối với máy chủ server qua kết nối Ethernet LAN/3G để cùng gửi dữ liệu về server (Hình 4). Server sẽ cùng xử lý các dữ liệu từ các client gửi về và trả lại kết quả sau khi xử lý cho các client để hiển thị thông tin giao dịch ra màn hình LCD.

2.1.2 Bảo mật và backup cho hệ thống

Phân tích rủi ro:

– Thẻ có thể bị đọc thông tin khi dùng các đầu đọc có tần số 13.56 MHz với chuẩn ISO, nên các thông tin ghi trên thẻ có thể không hoàn toàn được bảo mật.

– Thẻ có thể bị làm giả để thực hiện các thanh toán trong hệ thống.

– Cơ sở dữ kiệu trên máy tính có thể bị truy cập trái phép để thực hiện các thao tác thay đổi thông tin.

– Dữ liệu của hệ thống có thể bị mất khi ổ cứng bị hỏng hoặc mất điện đột ngột.

Khắc phục rủi ro:

– Không ghi bất kì thông tin gì lên thẻ, mọi thông tin tài khoản đều lưu trữ trên hệ thống để tránh thông tin người dùng bị sao chép.

– Mỗi một thẻ RFID hoặc NFC khi sản xuất ra luôn có một số ID riêng biệt và không trùng với bất kỳ thẻ nào theo quy định của công nghệ sản xuất thẻ vì thế người quản lý có thể kiểm tra những thẻ giả muốn thanh toán vào hệ thống qua ID đã được lưu trữ.

– Để tránh việc xâm nhập trái phép vào cơ sở dữ liệu của hệ thống, người dùng sẽ tạo tài khoản và mật khẩu đăng nhập, chỉ những người dùng có thông tin này mới có thể đăng nhập vào cơ sở dữ liệu được. Đồng thời có sự phân cấp giữ người dùng từ quản lý đến nhân viên.

– Để tránh việc mất dữ liệu của hệ thống khi mất điện hoặc hỏng ổ cứng, công nghệ backup dữ liệu RAID sẽ được áp dụng để ngăn ngừa việc mất dữ liệu.

2.2 Hệ thống thanh toán phí giữ xe bằng công nghệ RFID

2.2.1 Mục tiêu của hệ thống

– Thanh toán điện tử thay thế cho thanh toán tiền mặt.

– Thanh toán nhanh chóng, thuận tiện, tiết kiệm được thời gian, giảm tắc nghẽn tại điểm trả phí.

– Hệ thống chạy ổn định, an toàn và bảo mật.

– Hướng tới giải pháp quản lý các bãi giữ xe tập trung.

2.2.2 Các thành phần của hệ thống

Hệ thống thanh toán phí giữ xe bao gồm các phần: thẻ RFID, đầu đọc RFID, server, phần mềm xử lý và cơ sở dữ liệu.

Hình 5 trình bày một thẻ RFID dùng cho hệ thống thanh toán điện tử bãi giữ xe tự động trong bài viết.

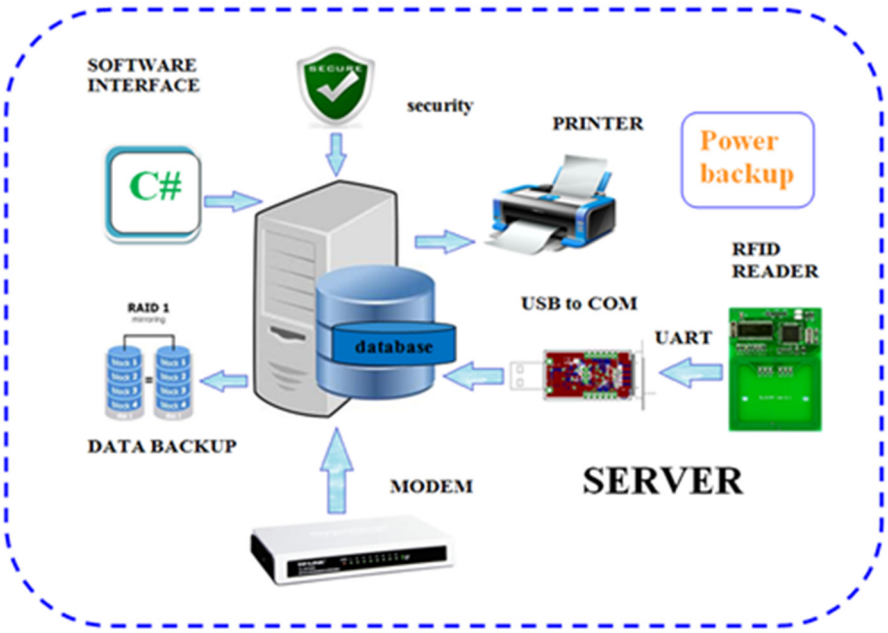
Hình 5: Hình dạng thẻ thanh toán điện tử RFID cho bãi giữ xe

Server hệ thống gồm các bộ phận sau: Máy tính có cấu hình cao chạy hệ điều hành Windows, cổng kết nối Ethernet, kết nối COM ảo, một đầu đọc RFID tần số 13.56 MHz chuẩn ISO 15693, một thiết bị chuyển đổi USB-COM, phần mềm xử lý

chạy trên server viết bằng ngôn ngữ C#, cơ sở dữ liệu thiết kế bằng MySQL, cơ chế bảo mật (security), hệ thống backup dữ liệu RAID và máy in để lịch sử thanh toán.

a. Mô hình cho server

Mô hình server trên Hình 6 là một máy chủ chứa hệ quản trị cơ sở dữ liệu MySQL cho hệ thống thanh toán điện tử. Người quản lý hệ thống có thể thao tác ghi/đọc thông tin vào cơ sở dữ liệu qua phần mềm xử lý viết bằng ngôn ngữ C#. Phần mềm này sẽ đọc ID thẻ từ đầu đọc RFID gửi về máy tính qua USB-COM. Trên giao diện phần mềm, người quản lý có thể thay đổi thông tin tài khoản và nạp tiền vào tài khoản thẻ, in xuất lịch sử thanh toán khi người dùng thẻ có yêu cầu. Để bảo mật cho dữ liệu trên server, người dùng sẽ được phân cấp tài khoản để tránh việc xâm nhập trái phép vào cơ sở dữ liệu. Sử dụng công nghệ backup dữ liệu RAID 1 để tránh trường hợp mất dữ liệu trên server khi ổ cứng bị hỏng hoặc do mất điện đột ngột. Công nghệ này sẽ giúp dữ liệu được an toàn và không làm gián đoạn hoạt động của hệ thống. Server sẽ nhận dữ liệu từ các client qua kết nối Ethernet, xử lý dữ liệu đó và trả lại kết quả cho client thông qua modem. USB-COM được dùng để giả lập cổng COM ảo để truyền dữ liệu theo chuẩn UART từ RFID reader lên máy tính.



Hình 6: Mô hình server cho hệ thống thanh toán điện tử

b. Phương thức xử lý dữ liệu và trừ tiền trong tài khoản

Tại client, ID của thẻ người dùng sẽ được gửi về server thông qua kết nối Ethernet. Server sẽ nhận ID và so sánh, kiểm tra xem ID có thuộc cơ sở dữ liệu hay không.

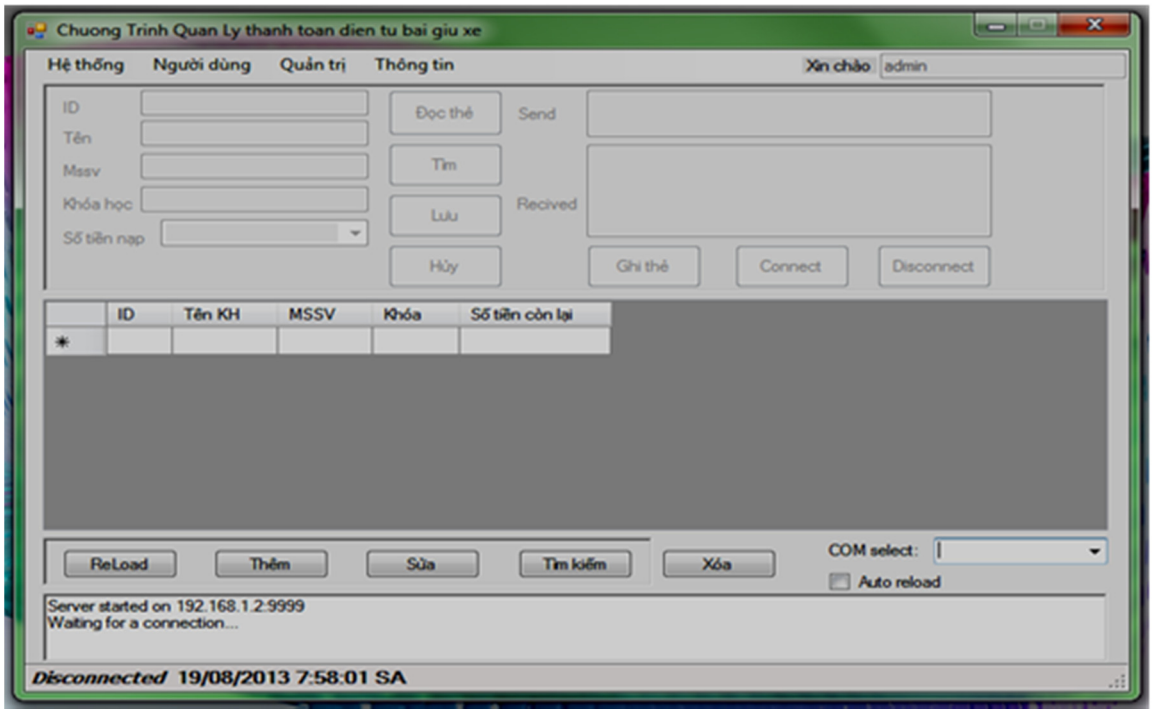
– Nếu ID có tồn tại trong hệ thống thì server sẽ xử lý tiếp ID gửi về thanh toán cho loại xe nào. Nếu là xe đạp thì trừ 500 đồng trong tài khoản của ID đó sau đó gửi trả về client thông báo về số tiền còn lại. Nếu loại thanh toán là xe máy thì server sẽ tiếp tục trừ tiền trong tài khoản của ID đó 1.000 đồng sau đó gửi thông báo về client số tiền còn lại.

– Nếu ID hết tiền trong tài khoản thì server sẽ gửi thông báo về client là ID đã hết tiền và không có quyền thanh toán.

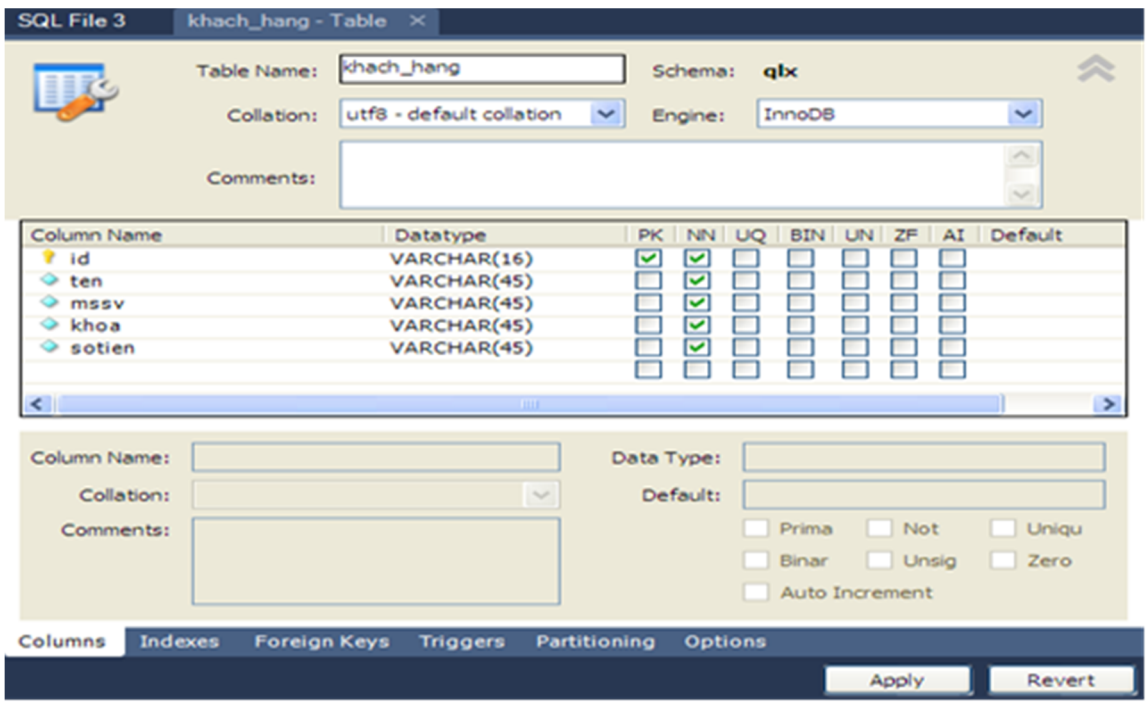
– Nếu ID không tồn tại trong hệ thống thì server sẽ gửi thông báo về client là ID thẻ này không thuộc hệ thống.

c. Phần mềm xử lý

Phần mềm xử lý "chương trình quản lý thanh toán điện tử phí giữ xe" được viết bằng ngôn ngữ C# trên nền Microsoft Visual Studio 2010 [10]. Có các chức năng: tự động detect cổng COM, đọc ID của thẻ RFID và lưu vào cơ sở dữ liệu, cấu hình kết nối từ client đến server, quản lý thông tin người dùng, lưu lịch sử giao dịch, xuất lịch sử thanh toán ra file Excel, tài khoản được phân quyền 2 cấp (admin và user), có chức năng là một server chính, nhận và xử lý thông tin giao dịch từ các client gửi về.



Hình 7: Giao diện phần mềm xử lý trên server



Hình 8: Hệ cơ sở quản trị dữ liệu MySQL

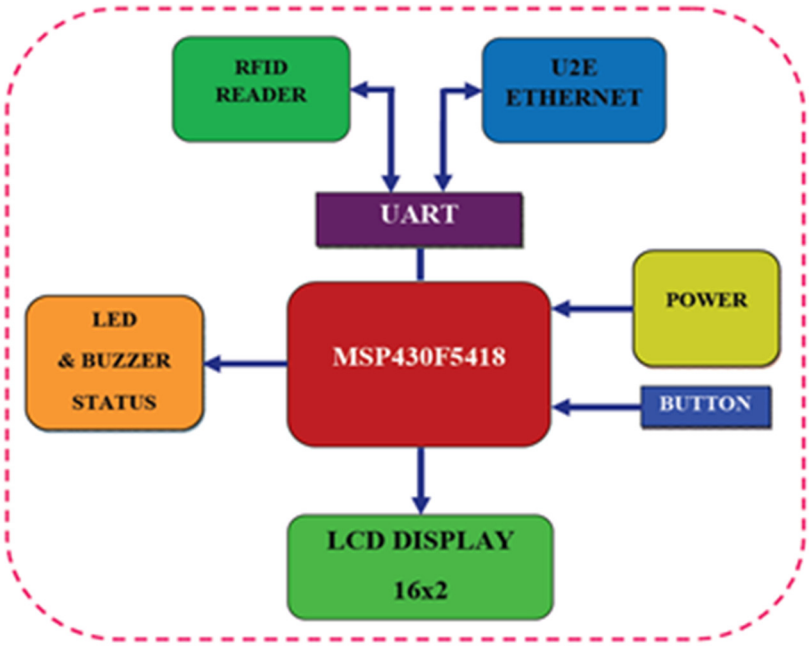
d. Cơ sở dữ liệu

Cơ sở dữ liệu được xây dựng trên hệ quản trị cơ sở dữ liệu MySQL mã nguồn mở của Orache (Hình 8).

e. Module Client

Sơ đồ khối của module client Hình 9

Phần cứng client sẽ bao gồm vi điều khiển MSP430F5418, đầu đọc RFID reader, module ethernet U2E, màn hình LCD text, mạch nguồn, nút nhấn, LED trạng thái và loa báo. Module thực hiện chức năng đọc ID thẻ và gửi về server, chờ nhận kết quả thanh toán từ server và hiển thị thông tin thanh toán ra LCD.



Hình 9: Sơ đồ khối module client

Nguyên tắc hoạt động của client

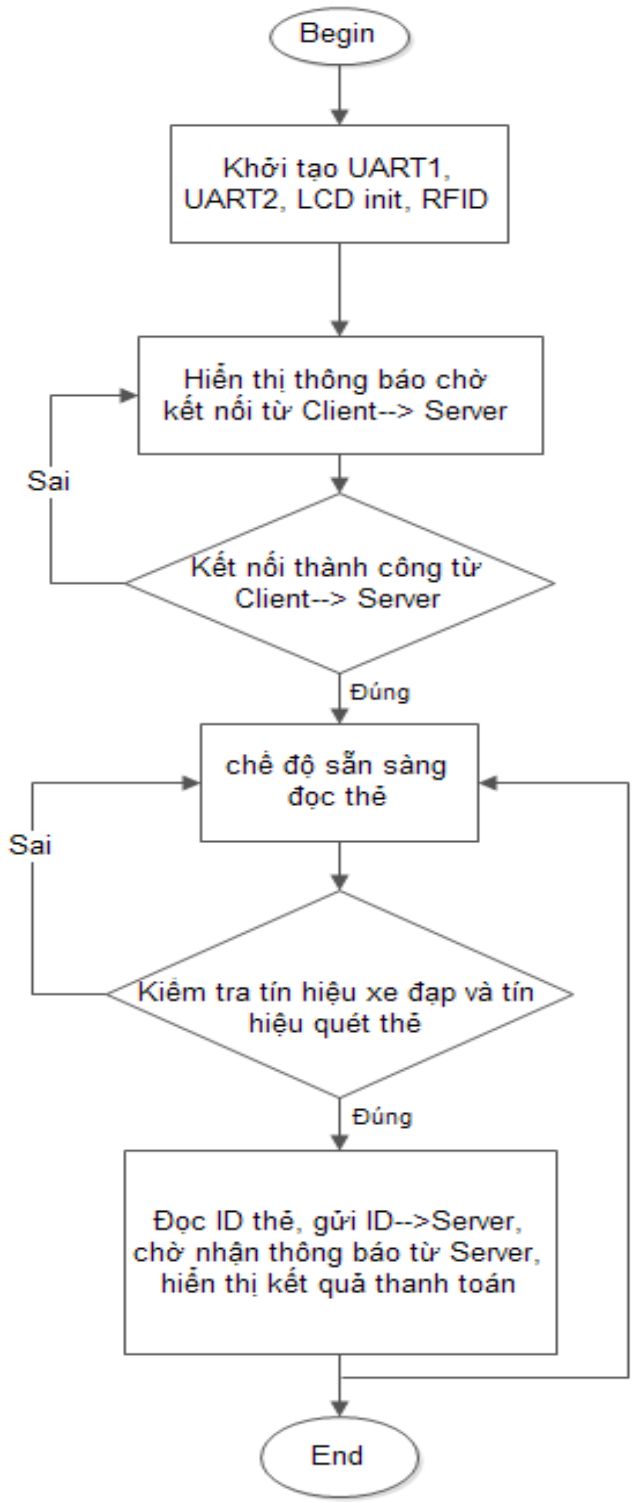
– Client có 2 chế độ, thanh toán cho xe máy và thanh toán cho xe đạp. Ở chế độ mặc định sẽ là thanh toán cho xe máy, tức là ID thẻ sẽ được đọc trực tiếp và gửi về vi điều khiển, trường hợp này là thanh toán xe máy. Nếu thanh toán cho xe đạp người gửi xe cần nhấn button sau đó ID của thẻ mới được đọc. Mục đích của việc trên là nhằm xác định giao dịch nào là thanh toán xe máy, giao dịch nào là thanh toán xe đạp.

– Thẻ RFID của người dùng được quét qua đầu đọc trang bị tại client. ID của thẻ sẽ được vi điều khiển MSP430F5418 nhận về từ RFID reader qua chuẩn giao tiếp UART. Đây là chế độ mặc định thanh toán cho xe máy. Khi đó, vi điều khiển MSP430F5418 sẽ gửi ID vừa nhận về server, server nhận được ID, server sẽ xử lý và trả về các trường hợp sau: Nếu tài khoản còn tiền thì sẽ trừ tiền và gửi về "số tiền còn lại", nếu tài khoản hết tiền thì sẽ gửi về thông báo "thẻ hết tiền", nếu ID thẻ không tồn tại sẽ gửi về thông báo "ID không thuộc hệ thống". Phía client sẽ nhận được các thông báo trên và sẽ cho hiển thị ra LCD để người thanh toán biết.

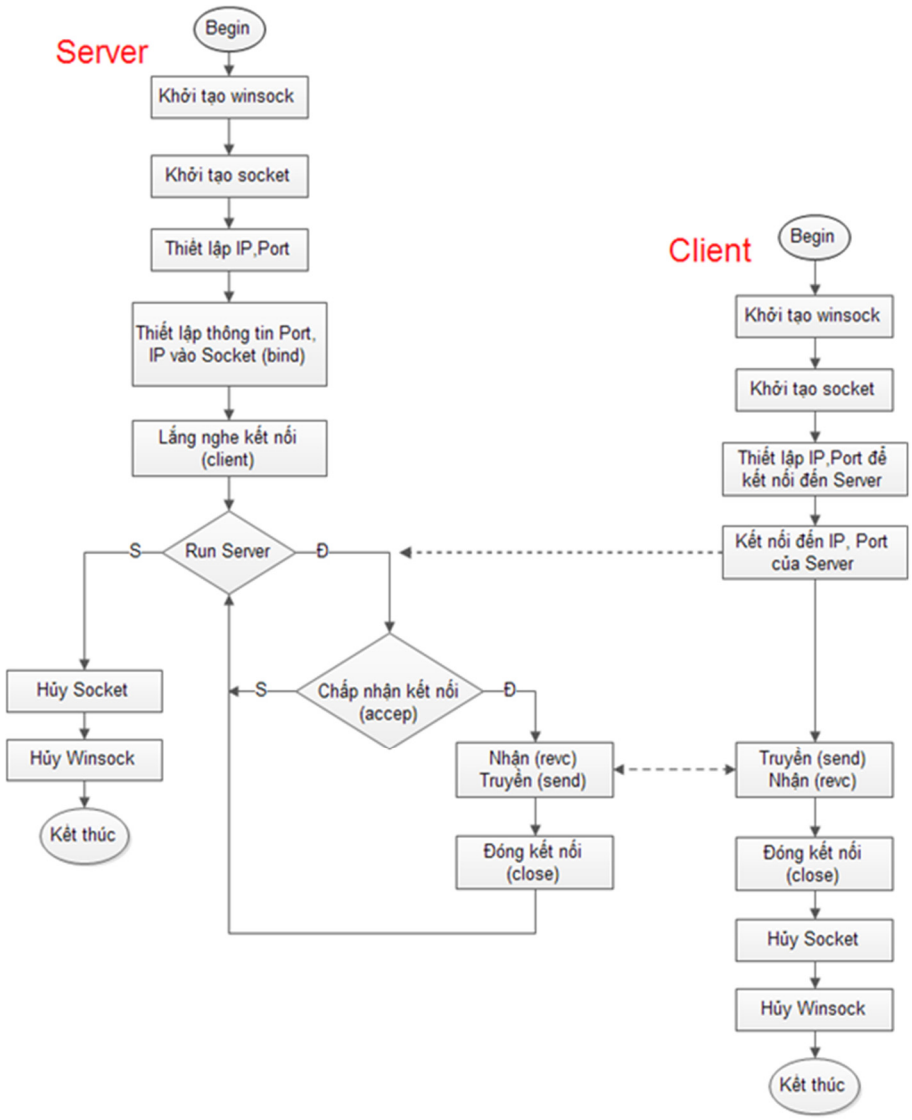
– Trường hợp thanh toán là xe đạp, người quản lý thanh toán sẽ nhấn một nút (button) trên module client và một đoạn mã sẽ được gửi về server nhằm mục đích thông báo cho server biết giao dịch sắp diễn ra là thanh toán phí cho xe đạp. Khi đó, phía server sẽ chờ đợi nhận tiếp ID từ client gửi về, tiếp tục server xử lý trừ tài khoản thanh toán cho xe đạp. Tương tự như xe đạp, khi thanh toán cho xe máy server cũng gửi về 3 trường hợp: số tiền còn lại, thẻ hết tiền, ID không tồn tại. Phía client nhận được và hiển thị thông báo thanh toán lên LCD tương ứng với các trường hợp trên. Hình 10 mô tả giải thuật chương trình của module client.

Kết nối giữa server và client: Theo Stream sockets dựa trên giao thức TCP (transmission Control Protocol), là giao thức hướng luồng (stream oriented). Việc truyền dữ liệu chỉ thực hiện giữa 2 tiến trình đã thiết lập kết nối. Giao thức này đảm bảo dữ liệu được truyền đến nơi nhận một cách đáng tin cậy, đúng thức tự nhờ vào cơ chế quản lý luồng lưu thông trên mạng và cơ chế chống tắc nghẽn. Hình 11 mô tả giải thuật chương trình kết nối giữa server và client qua giao thức TCP/IP.

Giải thuật cho client



Hình 10: Giải thuật cho module client



Hình 11: Sơ đồ giải thuật kết nối giữa server và client theo giao thức TCP/IP

Hình 12: Sơ đồ kết nối server + LAN + các client

Các client sẽ luôn kết nối về server để truyền dữ liệu cho server xử lý, trong một thời điểm có thể có nhiều client gửi dữ liệu về server cùng lúc. Server sẽ tự chia tiến trình để cùng lúc xử lý hết các dữ liệu gửi về và gửi trả lại kết quả cho client. Hình 12 là sơ đồ kết nối các client đến server.

f. Cơ chế hoạt động của hệ thống

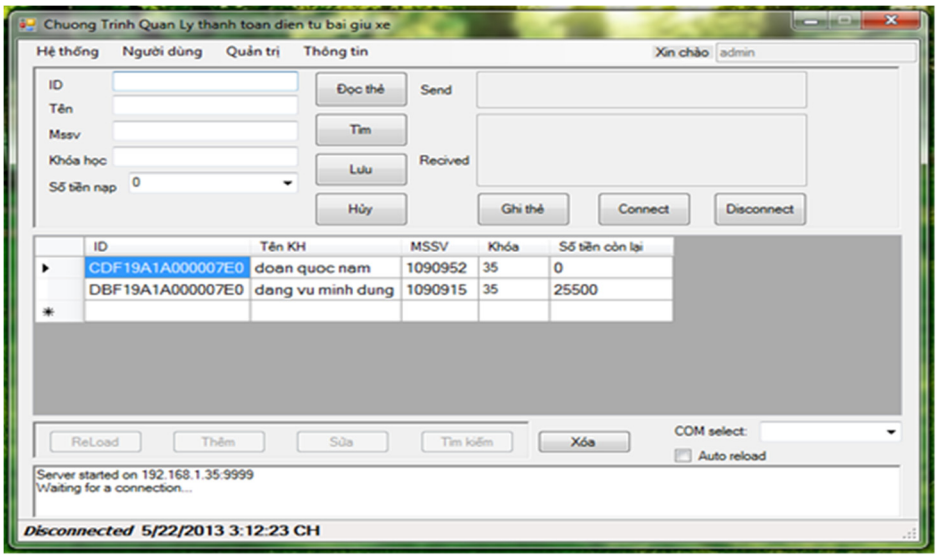
Hai bước thực hiện việc thanh toán điện tử là: thứ nhất người dùng đến quầy dịch vụ mua thẻ thanh toán điện tử RFID và nạp tiền vào tài khoản. Thứ hai là gửi xe và thanh toán.

– Đến quầy dịch vụ mua thẻ thanh toán: khi người sử dụng muốn dùng hệ thống thanh toán điện tử, trước hết họ phải đến quầy dịch vụ để mua thẻ thanh toán, cung cấp thông tin cá nhân theo quy định để lưu vào cơ sở dữ liệu của hệ thống và sẽ nạp số tiền mà họ cần theo các mệnh giá quy định trước.

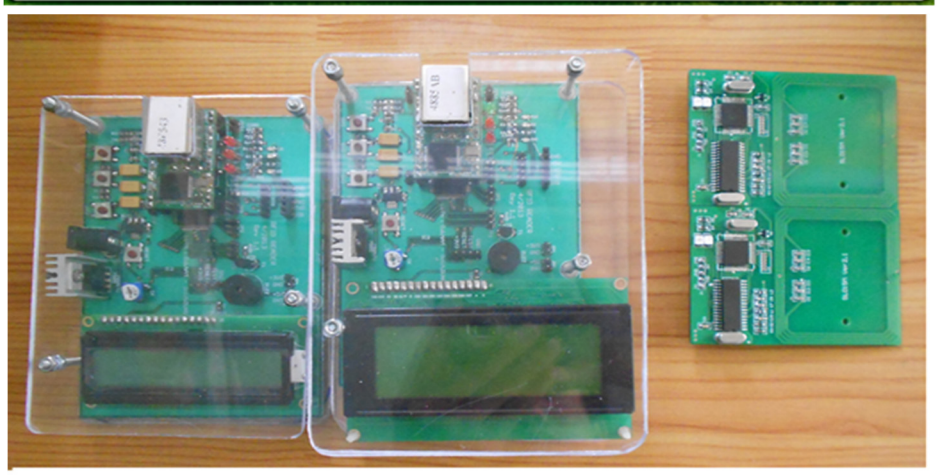
– Gửi xe và thanh toán: khi đến bãi giữ xe, người dùng thẻ thanh toán sẽ gửi xe vào bãi. Sau đó, khi lấy xe ra khỏi bãi xe thì người dùng cần quét thẻ tại đầu đọc RFID của client để thanh toán. Nếu giao dịch được thông báo thành công thì người dùng đã thanh toán xong phí giữ xe và có thể lấy xe ra khỏi bãi. Tất nhiên là khâu quản lý xe sẽ thuộc trách nhiệm của chủ giữ xe, không nằm trong mục tiêu của hệ thống này.

2.3 Kết quả đạt được

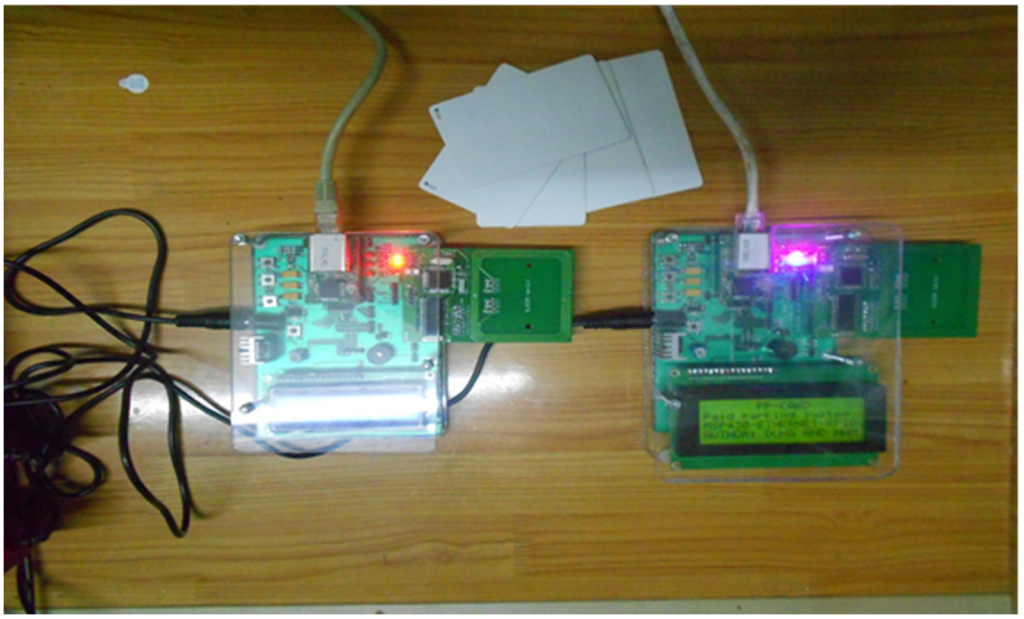
Một hệ thống thanh toán phí giữ xe bằng công nghệ RFID đã được thiết kế, chế tạo, thử nghiệm thành công với 2 đầu đọc thẻ và 01 máy tính làm chức năng server quản lý. Các mạch điện của hệ thống và giao diện của phần mềm quản lý trên server được trình bày ở các Hình 13, Hình 14 và Hình 15. Qua kiểm nghiệm thực tế, hệ thống đáp ứng được các yêu cầu đã đặt ra ban đầu.



Hình 13: Phần mềm quản lý trên máy tính



Hình 14: Client module



Hình 15: Hai client module cùng kết nối về server

3 KẾT LUẬN

Nội dung bài viết đã trình bày việc thiết kế, chế tạo thử nghiệm một hệ thống thanh toán điện tử bằng công nghệ RFID và NFC và được minh họa cụ thể bằng ví dụ thanh toán chi phí giữ xe. Đây là một giải pháp ứng dụng có tính khả thi cao và có thể áp dụng cho các dịch vụ công cộng như thanh toán phí giữ xe tại các điểm giữ xe trong Trường Đại học Cần Thơ.

Hướng nghiên cứu tiếp theo mà nhóm tác giả tiếp tục thực hiện là xây dựng hệ thống thanh toán không sử dụng kết nối server-client, tức là các client sẽ độc lập thực hiện việc quét thẻ và thanh toán. Hệ thống thanh toán client độc lập đòi hỏi chế độ bảo mật rất cao và khi đó đầu đọc và thẻ RFID cũng cần có chức năng bảo mật. Với đặc điểm này, hệ thống thanh toán client độc lập không những có thể áp dụng được cho các điểm giữ xe có thu phí mà còn có thể áp dụng trong thanh toán vé xe buýt, mua hàng tại siêu thị, quán ăn, khách sạn và các ứng dụng thanh toán khác.

TÀI LIỆU THAM KHẢO

1. Near field communication. URL: http://en.wikipedia.org/wiki/Near_field_communication.
2. FeliCa by Sony. URL: http://newconstruct.wordpress.com/2008/03/27/36/
3. Klaus Finkenzeller, and Dorte Muller, *RFID Handbook*, 3rd edition, Wiley, August 2010.
4. RFID basics. URL: http://www.priority1design.com.au/rfid_design.html
5. EZ-LINK. URL: http://www.ezlink.com.sg
6. Octopus. URL: http://www.octopus.com.hk
7. NTT Docomo. URL: http://www.nttdocomo.co.jp
8. Tuoitre Online, "*2013: dân số Việt Nam đạt 90 triệu người*", đăng ngày 26/12/2012. URL: http://tuoitre.vn/chinh-tri-xa-hoi/song-khoe/526788/2013-dan-so-viet-nam-dat-90-trieu-nguoi.html
9. Mai Tiểu Long, *RFID – Từ Khái Quát Đến tập lệnh*, RFID Team, Texas Instruments.
10. Simon Robinson, *et al*., *Professional C#,* 2nd Edition, Wrox Press, 2002.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

TỔNG QUAN VỀ AN NINH TRÊN ĐIỆN TOÁN ĐÁM MÂY

Trần Cao Đệ[1]
[1] Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Overview on cloud computing security

Từ khóa:
An toàn mạng, kiến trúc hướng dịch vụ, IaaS, PaaS, SaaS

Keywords:
Network security, Service oriented architecture, IaaS, PaaS, SaaS

ABSTRACT

In recent years cloud computing has emerged as a new development phase of the Internet. Cloud computing, its service models in particular, allows the use of hardware, software as services, that leads a fundamental change in the application of information technology in practice - shift from investment to the hire of computing resources. However, cloud adoption is still controversy issues surrounding the safety and security of system. This paper provides an overview on safety issues and security of cloud computing, from the perspective of architecture, services as well as main characters of cloud computing.

TÓM TẮT

Mấy năm gần đây điện toán đám mây (cloud computing) nổi lên như là một giai đoạn phát triển mới của Internet. Điện toán đám mây mà cụ thể là các mô hình dịch vụ của nó cho phép sử dụng phần cứng, phần mềm như là dịch vụ làm thay đổi căn bản việc ứng dụng công nghệ thông tin trong thực tiễn – chuyển từ đầu tư sang thuê bao. Tuy nhiên, việc chấp nhận điện toán đám mây vẫn còn khá dè dặt xoay quanh vấn đề an toàn và an ninh hệ thống. Bài viết này cung cấp một cái nhìn tổng quan về vấn đề an toàn và an ninh của điện toán đám mây, từ góc độ kiến trúc cho đến dịch vụ cũng như tính chất của đám mây điện tử.

1 GIỚI THI ỆU

Điện toán đám mây đang nhanh chóng nổi lên như một xu hướng công nghệ, hầu hết các nhà cung cấp công nghệ hoặc sử dụng phần mềm, phần cứng và cơ sở hạ tầng đều có thể tận dụng. Công nghệ và kiến trúc dịch vụ điện toán đám mây cho phép khách hàng của các dịch vụ này không sở hữu tài sản trong các đám mây điện tử nhưng trả tiền trên cơ sở mỗi lần sử dụng. Về bản chất, họ đang thuê cơ sở hạ tầng vật lý và các ứng dụng trong kiến trúc dùng chung. Dịch vụ điện toán đám mây có thể từ lưu trữ dữ liệu cho đến các ứng dụng Web của người dùng cuối, cùng với các dịch vụ điện toán tập trung khác.

Một sự khác biệt quan trọng giữa mô hình tính toán truyền thống và điện toán đám mây là khả năng mở rộng và tính chất co giãn mà điện toán đám mây cung cấp. Thay vì một kiến trúc hệ thống tĩnh, điện toán đám mây hỗ trợ khả năng tự động mở rộng quy mô và nhanh chóng thu hẹp quy mô, cung cấp cho khách hàng các dịch vụ có độ tin cậy cao, thời gian đáp ứng nhanh chóng và sự linh hoạt để xử lý các biến động thông lượng và nhu cầu. Điện toán đám mây cũng hỗ trợ nhiều người dùng, cung cấp các hệ thống cấu hình trong suốt một cách thống nhất mang tính chia sẻ được. Công nghệ ảo hóa cho phép các nhà cung cấp điện toán đám mây để chuyển đổi một máy chủ thành nhiều máy ảo, do đó loại bỏ máy tính client-server với các hệ thống một mục đích. Điều này tối đa hóa khả năng phần cứng và cho phép khách hàng tận dụng triệt để và kinh tế theo quy mô sử dụng.

Theo nghiên cứu của Gartner đưa ra trong 4, doanh thu trên toàn thế giới từ điện toán đám mây năm 2009 là 58.6 tỷ USD, dự báo đạt 148 tỷ USD vào 2015. Nghiên cứu gần đây của IDC 4 chỉ rõ doanh thu trên toàn thế giới từ điện toán đám mây công cộng (public cloud) năm 2010 là hơn 21,5 tỷ USD và sẽ đạt 72,9 tỷ USD vào năm 2015, tốc độ tăng trưởng hàng năm là 27,6%. Tốc độ tăng trưởng này gấp hơn bốn lần so với tốc độ tăng trưởng dự báo cho toàn thị trường CNTT trên thế giới nói chung (6,7%). Cũng theo IDC, điện toán đám mây luôn nằm trong tốp 10 các công nghệ đình đám nhất trong các năm từ 2009 đến nay trong lĩnh vực công nghệ thông tin.

Mặc dù có những lợi ích và tiềm năng phát triển như vậy nhưng điện toán đám mây vẫn còn gặp nhiều dè dặt trong chấp nhận và nhân rộng chủ yếu là vấn đề an toàn. Tạm gác qua vấn đề kỹ thuật, từ quan điểm của người dùng thì các vấn đề chủ yếu cản trở việc chấp nhận điện toán đám mây là :

– Các doanh nghiệp giao việc quản lý an ninh cho một bên thứ ba nên có thể không kiểm soát được vấn đề an ninh.

– Tài sản của nhiều người thuê bao khác nhau nằm trong cùng một vị trí (server) và sử dụng cùng một dịch vụ mà không biết rõ cách thức kiểm soát an ninh, bảo mật của nhà cung cấp.

– Thiếu sự đảm bảo an toàn, an ninh trong các hợp đồng giữa người dùng (người thuê bao) điện toán đám mây và các nhà cung cấp điện toán đám mây.

– Các máy chủ tập trung các tài sản có giá trị và công khai cơ sở hạ tầng làm tăng xác suất các cuộc tấn công nguy hiểm.

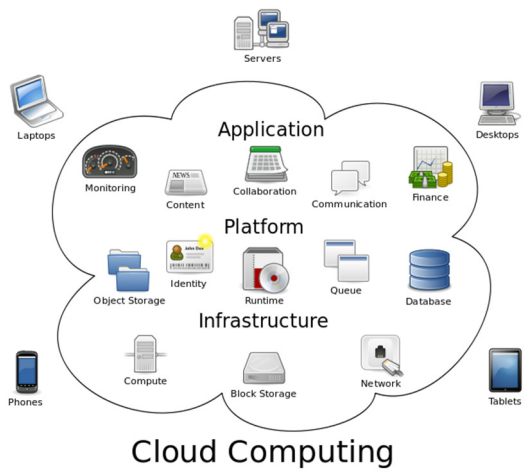
Bài viết này sẽ phân tích những thách thức và các vấn đề liên quan an ninh trên điện toán đám mây, những điểm yếu trong mô hình điện toán đám mây. Những vấn đề này sẽ được xem xét từ góc độ kiến trúc dịch vụ đến mô hình cùng với các đặc trưng của điện toán đám mây.

2 ĐIỆN TOÁN ĐÁM MÂY

2.1 Khái niệm

Thuật ngữ "Điện toán đám mây" (cloud computing) ra đời vào khoảng giữa năm 2007 nhằm để khái quát lại các hướng phát triển mới của công nghệ thông tin nhờ vào mạng Internet băng thông rộng và các trung tâm điện toán khổng lồ của các hãng công nghệ như Google, Amazon, IBM, Microsoft,… Điện toán đám mây gắn liền với một quan niệm mới về công nghệ thông tin, đó là: các nguồn lực điện toán khổng lồ như phần mềm, dữ liệu dịch vụ sẽ nằm tại các máy chủ ảo (đám mây) trên Internet thay vì trong máy tính của tổ chức, cá nhân để mọi người kết nối và sử dụng khi cần. Với các dịch vụ hạ tầng, phần mềm sẵn có trên Internet, doanh nghiệp không phải mua và duy trì hàng trăm, thậm chí hàng nghìn máy tính cũng như phần mềm cho công ty. Họ có thể thuê toàn bộ hạ tầng công nghệ thông tin như thuê bao điện thoại hay sử dụng điện, nước hàng ngày.

Theo định nghĩa của NIST 1: "Điện toán đám mây là một mô hình cho phép thuận tiện, truy cập mạng theo yêu cầu đến một nơi chứa các nguồn tài nguyên tính toán có thể chia sẻ và cấu hình được (ví dụ: mạng, máy chủ, lưu trữ, ứng dụng và dịch vụ), ở đó chúng có thể được cung cấp và phát hành nhanh chóng với nỗ lực quản lý hoặc tương tác với nhà cung cấp tối thiểu".



Hình 1: Minh họa cho điện toán đám mây theo Wikipedia 2

Điện toán đám mây đôi khi còn được coi là thế hệ Internet mới. Tự điển mở Wikipedia 2 định nghĩa: "Điện toán đám mây là việc sử dụng các tài nguyên máy tính (phần cứng và phần mềm) có sẵn từ xa và truy cập được qua mạng (thường là Internet)". Hình 1 phát họa khái niệm về điện toán đám mây. Điện toán đám mây cung cấp các tiện ích để truy cập vào tài nguyên chia sẻ và cơ sở hạ tầng chung, cung cấp dịch vụ theo yêu cầu qua mạng để thực hiện các hoạt động đáp ứng nhu cầu tác nghiệp. Vị trí của nguồn lực vật chất và thiết bị được truy cập là trong suốt, không được biết (và cũng không cần biết) đối với người dùng cuối (end user). Nó cũng cung cấp phương tiện cho người sử dụng (hay khách hàng) để phát triển, triển khai và

quản lý các ứng dụng của họ trên các đám mây, kể cả ảo hóa các nguồn tài nguyên, tự bảo trì và quản lý các ứng dụng.

2.2 Các dịch vụ cơ bản của điện toán đám mây

Điện toán đám mây cung cấp 3 mô hình dịch vụ cơ bản: dịch vụ hạ tầng (IaaS), dịch vụ nền tảng (PaaS) và dịch vụ phần mềm (SaaS), với một số đặc trưng chính : thuê bao theo yêu cầu, nhiều thuê bao, dùng bao nhiêu trả bấy nhiêu. Về mặt kỹ thuật, đám mây là một tập hợp tài nguyên tính toán rộng lớn và cung cấp 3 dịch vụ nói trên.

Dịch vụ hạ tầng (IaaS - Infrastructure as a Service) nói đến khả năng cung cấp cho khách hàng tài nguyên phần cứng như khả năng tính toán, xử lí (tức là CPU), lưu trữ, mạng và các tài nguyên khác, để khách hàng có thể triển khai và chạy phần mềm tùy ý, trong đó có thể bao gồm các hệ điều hành và các ứng dụng. Khách hàng sẽ không quản lý hoặc kiểm soát các cơ sở hạ tầng điện toán đám mây nhưng có kiểm soát đối với các hệ điều hành, lưu trữ, ứng dụng triển khai và kiểm soát có hạn chế các thành phần mạng được cung cấp. IaaS gắn liền với việc ảo hóa tài nguyên phần cứng. Những Nhà cung cấp IaaS điển hình là Amazon EC2, GoGrid và HP [19]. Nhà cung cấp dịch vụ Iaas sẽ chịu trách nhiệm mọi công việc nặng nề về thiết lập hạ tầng, thiết lập chức năng để cung cấp hạ tầng và thu phí thuê bao hạ tầng. Khách hàng có thể tăng giảm một cách tự động, thuận tiện về tài nguyên họ cần. Một số lượng lớn tài nguyên có thể được cung cấp và sẵn dùng trong một thời gian ngắn sau khi được yêu cầu và quan trọng hơn là hành vi của hệ thống không thay đổi, không có lỗi tiềm tàng do chuyển đổi từ hệ thống nhỏ sang hệ thống lớn hơn hay ngược lại.

Dịch vụ nền tảng (PaaS- Platform as a Service): đó là khả năng cung cấp cho khách hàng nền tảng để triển khai trên cơ sở hạ tầng điện toán đám mây các ứng dụng do khách hàng tạo ra từ ngôn ngữ lập trình và các công cụ hỗ trợ của nhà cung cấp. Khách hàng không quản lý hoặc kiểm soát cơ sở hạ tầng điện toán đám mây cơ bản như mạng, máy chủ, hệ điều hành, thiết bị lưu trữ, nhưng có kiểm soát đối với các ứng dụng triển khai và có thể thực hiện cấu hình môi trường lưu trữ. Có thể coi dịch vụ này cung cấp các phần mềm hệ thống cần thiết như là ngôn ngữ lập trình, môi trường lập trình, môi trường thực thi, hệ điều hành để người dùng truy cập tài nguyên và tạo ra các ứng dụng của mình. Các nhà cung cấp dịch vụ này điển hình như Microsoft Windows Azure, Google App Engine.

Dịch vụ phần mềm (SaaS – Software as a Service) đó là khả năng cung cấp cho khách hàng sử dụng các ứng dụng (phần mềm) của nhà cung cấp đang chạy trên một cơ sở hạ tầng điện toán đám mây. Các ứng dụng có thể truy cập từ các thiết bị khác nhau thông qua một giao diện người dùng như một trình duyệt web. Khách hàng không quản lý hoặc kiểm soát các cơ sở hạ tầng cơ bản đám mây nhưng có thể thiết lập cấu hình cho ứng dụng phù hợp với mình. Nhiều người trong chúng ta chắc đã sử dụng phần mềm trên điện toán đám mây của Google như: Gmail, Google Docs, trình tìm kiếm của Google,... Đó là những ví dụ điển hình về SaaS. Dịch vụ phần mềm được cung cấp dựa theo cơ chế dịch vụ web (web service) và các cổng thông tin điện tử (portal).

2.3 Các mô hình triển khai điện toán đám mây

Có bốn mô hình chính để triển khai điện toán đám mây đó là:

Hình 2: Năm đặc trưng quan trọng của điện toán đám mây

Đám mây công cộng (Public Cloud): đám mây được thiết lập và cung cấp cho rộng rãi người dùng thông qua Internet. Nó còn được biết như là đám mây nhiều thuê bao với các đặc trưng cơ bản là hạ tầng thống nhất, chính sách chung, nguồn lực chia sẻ cho nhiều thuê bao, đa qui mô. Mô hình đám mây này thường ít an toàn hơn các mô hình khác và thường chỉ cung cấp các dịch vụ phần mềm chung nhất như bộ phần mềm văn phòng, chat, họp trực tuyến,…

Đám mây riêng (Private Cloud): đám mây được thiết lập chỉ cho một tổ chức tương tự như một mạng nội bộ. Nó có thể được quản lý bởi chính tổ chức đó hoặc một bên thứ ba và có thể tồn tại trên cơ sở hạ tầng trước đó đã có. Mô hình này còn được gọi là đám mây nội bộ và thường chỉ dành quyền truy cập vào tài nguyên của nó cho người dùng trong nội bộ tổ chức là chủ sở hữu đám mây. Đặc điểm cơ bản của đám mây riêng là hạ tầng không đồng nhất, chính sách "may đo" và tùy chỉnh, tài nguyên dành riêng, cơ sở hạ tầng "cây nhà lá vườn". Tuy nhiên do chỉ có các tổ chức và người dùng được phép mới có thể truy cập nên nó có thể được bảo vệ bởi các quy trình, quy chế bảo mật riêng, điều này làm cho nó khó bị tấn công hơn.

Đám mây cộng đồng (Community Cloud): đây là dạng đám mây mà hạ tầng được chia sẻ bởi một vài tổ chức. Nó hỗ trợ một vài thứ chung của cộng đồng đó, chẳng hạn như nhiệm vụ, chính sách bảo mật, các chuẩn,…

Đám mây hỗn hợp (Hybrid Cloud): hạ tầng của đám mây là một sự kết nối của nhiều mô hình triển khai đám mây (chung, riêng, cộng đồng)

2.4 Đặc trưng của điện toán đám mây và lợi ích của nó

Điện toán đám mây có 5 đặc trưng quan trọng (xem Hình 2):

Khả năng co giãn (scalability): đám mây có thể cung cấp tài nguyên tính toán theo yêu cầu. Việc cung cấp này trên nguyên tắc là động và nhiều thuê bao, vì vậy tránh lãng phí (dùng bao nhiêu trả bấy nhiêu).

– Khả năng quản trị và vận hành (manageability): đây là khả năng điều khiển, kiểm soát hệ thống và tính cước phí thuê bao.

– Khả năng truy cập (accessibility) và khả chuyển: truy cập mọi lúc mọi nơi một cách nhất quán và khả năng truy cập với các thiết bị nhỏ, yếu (thin client) như là điện thoại di động.

– Hiệu năng cao và tối ưu hóa (performance and Optimization): hạ tầng đám mây giải quyết và che dấu mọi vấn đề phức tạp trong tính toán song song, cân bằng tải, lập lịch để cung cấp khả năng tính toán hiệu năng cao và tối ưu hóa.

– Khả năng sẵn dùng với độ tin cậy cao (availability): hạ tầng đám mây cùng được cung cấp rộng rãi cho người dùng với khả năng chịu đựng lỗi cao, hệ thống tồn tại lâu dài và khả năng bảo mật tốt.

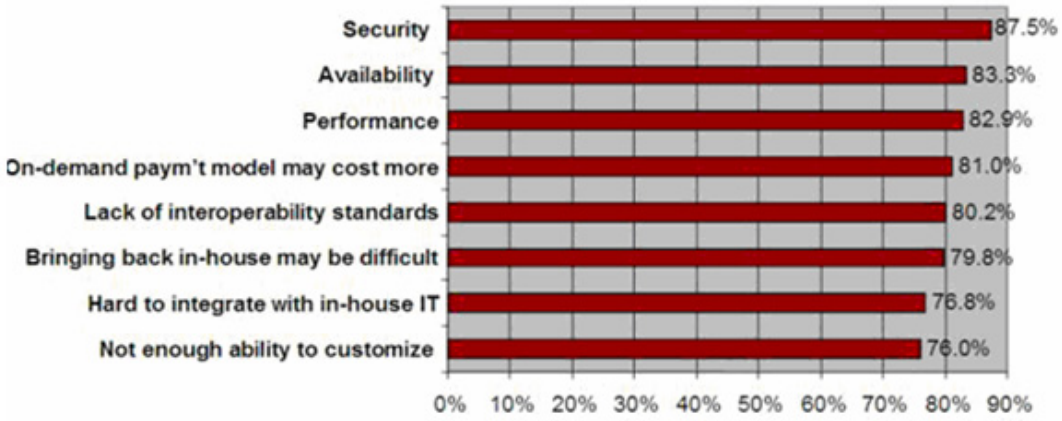
3 NGHIÊN CỨU CÓ LIÊN QUAN VỀ AN TOÀN TRÊN ĐIỆN TOÁN ĐÁM MÂY

Rõ ràng rằng có nhiều vấn đề cần bàn trong quyết định chấp nhận điện toán đám mây. Hình 3 cho một số vấn đề chính trên điện toán đám mây. Đây là những chủ đề luôn được quan tâm, theo nghiên cứu đã chỉ ra trong 5, vấn đề an toàn được đặt hàng đầu. Vấn đề thì khó nhưng lý do có vẻ thật đơn giản, bởi vì người ta đặt dữ liệu của mình (loại tài sản (rất) có giá trị) trên một máy tính ở "trên mây", dùng một phần mềm của ai đó (một nhà cung cấp) ở một nơi nào đó không biết rõ và chạy trên một máy ảo mà máy thật (tức là CPU vật lý) nằm ở đâu cũng không rõ nốt. Có rất nhiều nỗi lo lắng trong hoàn cảnh như vậy: mất dữ liệu, thông tin bị rò rỉ, bị theo dõi, bị lấy cắp, lừa đảo, nghẽn mạng,…

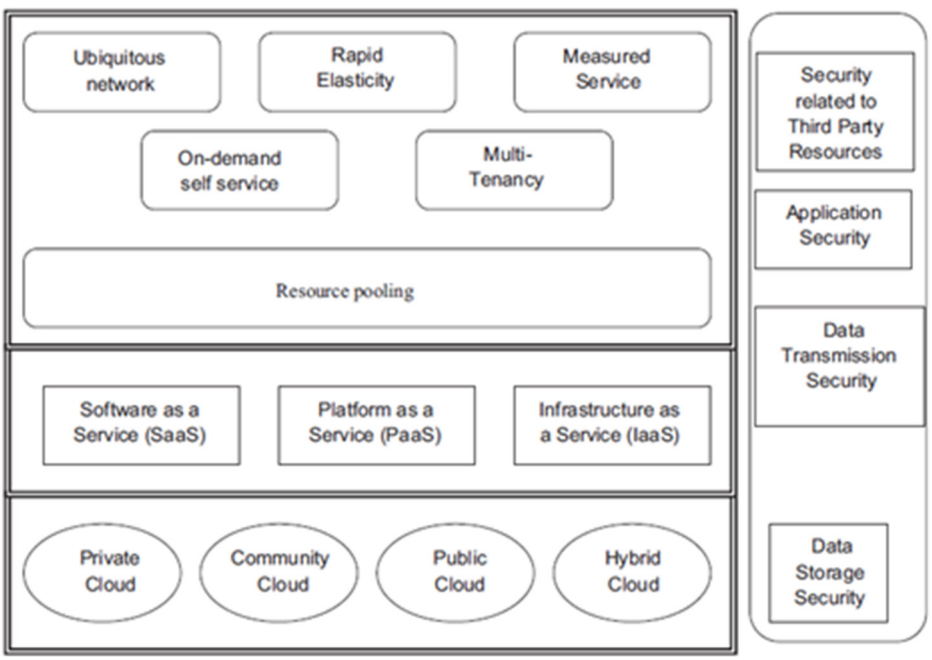
Nhiều nghiên cứu về vấn an toàn trên điện toán đám mây đã được thực hiện. Nhóm "Cloud Computing Use Cases" 6 thảo luận nhiều vấn đề liên quan tới mô hình điện toán đám mây trên các quan điểm của người dùng, nhà phát triển và kỹ sư bảo mật. Văn phòng an toàn thông tin và mạng Châu Âu (ENISA) 7 đưa ra các rủi ro, các ảnh hưởng, các điểm yếu của điện toán đám mây. Một nghiên cứu về an toàn được ràng buộc bằng hợp đồng 9 cố gắng đưa ra các đặc tả liên quan tới vị trí đặt dữ liệu, phân nhóm dữ liệu và phục hồi dữ liệu. Một số vấn đề chuyên biệt về an toàn trên điện toán đám mây như toàn vẹn dữ liệu, tính riêng tư và thông tin nhạy cảm được thảo luận trong 10. Vấn đề kỹ thuật tấn công, chẳng hạn kiểu tấn công XML (XML-attack) có thể tìm thấy trong 11, lỗ hổng bảo mật đám mây có thể tìm thấy trong 12. Các thách thức an toàn và bảo mật liên quan đến các mô hình dịch vụ của điện toán đám mây, đặc biệt là mô hình SaaS có thể tham khảo trong 13.

Vấn đề an toàn trên điện toán đám mây là rất phức tạp vì nó liên quan đến nhiều khía cạnh và chủ thể: kiến trúc, dịch vụ điện toán đám mây, đặc trưng của đám mây, thuê bao, chủ sở hữu, chính sách bảo

mật dữ liệu,… Hình 4 cho một hình ảnh rộng, khái quát về sự phức tạp của vấn đề này. Phần tiếp theo của bài viết sẽ đi vào tổng hợp về vấn đề an ninh trên điện toán đám mây từ góc độ kiến trúc, dịch vụ và một số đặc điểm chính của điện toán đám mây.



Hình 3: Các vấn đề quan tâm hàng đầu trong điện toán đám mây 5



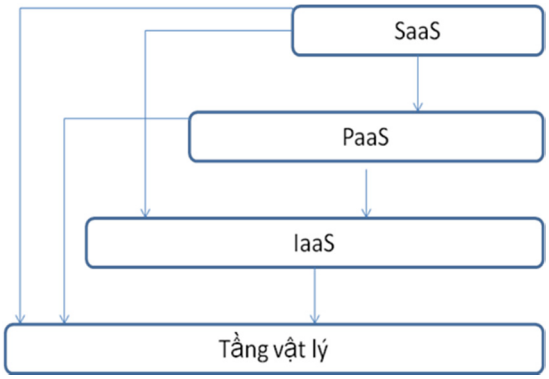
Hình 4: Sự phức tạp của vấn đề an toàn trên điện toán đám mây 13

4 MỘT SỐ VẤN ĐỀ ĐƯỢC QUAN TÂM VỀ AN TOÀN TRÊN ĐIỆN TOÁN ĐÁM MÂY

4.1 Vấn đề an toàn liên quan đến kiến trúc của điện toán đám mây

Một đám mây điện tử là một cụm máy tính kết nối nhau thông qua mạng cục bộ hoặc mạng diện rộng trên cơ sở ảo hóa tài nguyên phần cứng nhờ chức năng ảo hóa để cung cấp một cách trong suốt 3 dịch vụ cơ bản của điện toán đám mây là SaaS, PaaS và IaaS. Mô hình triển khai đám mây có thể là công cộng, đám mây riêng hoặc cộng đồng hay hỗn hợp như đã nói ở phần trên.

Các dịch vụ điện toán đám mây có kiến trúc phân tầng (layer), mỗi tầng cung cấp các dịch vụ và tiện ích (chức năng) riêng của nó trên cơ sở các dịch vụ và tiện ích của tầng thấp hơn (xem Hình 5). Vì vậy an ninh của hệ thống phụ thuộc vào an ninh của mỗi tầng được thiết kế và cài đặt kèm theo như là 1 dịch vụ hay tiện ích.



Hình 5: Kiến trúc phân tầng dịch vụ trong điện toán đám mây

4.1.1 An ninh ở mức hạ tầng

An ninh của các dịch vụ ở tầng thấp như tầng vật lý hay hạ tầng (IaaS) phụ thuộc vào nhà cung cấp, tức là chủ sở hữu của đám mây. Hiện tại, có một số nhà cung cấp dịch vụ IaaS nhưng chưa có chuẩn nào về an ninh cho các dịch vụ này. Về mặt nguyên tắc, khách hàng thuê bao dịch vụ IaaS có thể áp đặt các chính sách an ninh của mình bằng cách phát triển các dịch vụ hay tiện ích riêng thông qua các dịch vụ của tầng vật lý và các dịch vụ IaaS của nhà cung cấp. Chính sách về an toàn ở mức này là rất phức tạp vì nhiều chính sách khác nhau áp đặt lên cùng một môi trường phần cứng (vật lý).

Những mối đe dọa an toàn ở mức này có thể liên quan tới máy chủ ảo (Virtual Machine) như là vi-rút và các phần mềm độc hại khác. Nhà cung cấp dịch vụ chịu trách nhiệm chính về giải pháp cho vấn đề này. Khách hàng thuê bao cũng có thể thực hiện các giải pháp và chính sách an toàn riêng cho mình, từ đó làm gia tăng gánh nặng lên phần cứng và hiệu năng chung của hệ thống. Các máy chủ ảo vẫn có thể bị lây nhiễm hay bị kiểm soát bởi phần mềm độc hại. Trong trường hợp này, các chính sách an ninh của khách hàng có thể bị vô hiệu, như vậy nhà cung cấp dịch vụ phải là người có vai trò chính trong an ninh ở mức này. Ngoài ra, vì IaaS khai thác hạ tầng vật lý và chính sách chung như DNS Server, Switch, IP protocol,… Vì vậy, khả năng bị tấn công vào "khách hàng yếu nhất" sau đó "lây lan" cho các khách hàng khác. Vấn đề này hiện nay khách hàng thuê bao không thể can thiệp gì vì nhiều máy chủ ảo chia sẻ cùng tài nguyên vật lý như CPU, bộ nhớ, đĩa,… Mọi ánh xạ vật lý-máy ảo, máy ảo-vật lý đều thông qua một "bộ ảo hóa", nếu bộ này bị phần mềm độc hại kiểm soát thì toàn bộ khách hàng trong đám mây sẽ bị cùng một mối hiểm họa như nhau.

4.1.2 An ninh ở mức dịch vụ nền tảng

Ở mức trung gian, dịch vụ nền tảng (PaaS) dựa trên dịch vụ tầng dưới (IaaS) và cung cấp dịch vụ của mình cho tầng trên nó (SaaS). Ở mức này, các dịch vụ hay tiện ích về an toàn có thể được cài đặt thêm hoặc cấu hình các dịch vụ được cung cấp từ tầng dưới. Ở đây, người dùng có thể quản trị phần thuê bao của mình để tạo ra môi trường thực thi các ứng dụng. Hiện nay, dịch vụ PaaS của đám mây dựa trên mô hình kiến trúc hướng dịch vụ (SOA) vì vậy những nguy cơ về an toàn giống hệt như những nguy cơ an toàn của SOA như tấn công từ chối dịch vụ, tấn công XML và nhiều cách tấn công khác 1114.

Vì dịch vụ nền tảng là dịch vụ đa thuê bao, nhiều người dùng nên cơ chế xác thực, chứng thực là rất quan trọng. Trách nhiệm bảo mật và an toàn trong trường hợp này liên quan đến cả nhà cung cấp, người thuê bao và người dùng (user). Các dịch vụ PaaS phải cung cấp môi trường để phát triển ứng dụng bao gồm chức năng tác nghiệp, các chức năng an toàn và quản lí hệ thống. Nhà cung cấp cần có cơ chế bắt buộc chứng thực để truy cập các dịch vụ PaaS, người thuê bao có trách nhiệm phát triển hay cung cấp các chức năng bảo mật cần thiết thông qua cơ chế chứng thực chung và người dùng phải có trách nhiệm bảo vệ tài khoản đăng nhập cá nhân của mình.

4.1.3 An ninh ở mức dịch vụ phần mềm

Ở mức dịch vụ phần mềm (SaaS), các phần mềm được cung cấp như là dịch vụ trên mạng, sử dụng các chính sách bảo mật dữ liệu và tài nguyên khác từ các tầng bên dưới cung cấp. Một số dịch vụ phần mềm khá phổ biến hiện nay là Google Search Engine, Google mail… Khách hàng của các dịch vụ này không biết được dữ liệu của mình được quản lí và khai thác như thế nào và nó nằm ở đâu trên thế giới này. Vấn đề an ninh ở đây liên quan đến bảo mật dữ liệu, rò rỉ thông tin nhạy cảm và nguy cơ bị tấn công từ chối truy cập… Trách nhiệm về an toàn được chia sẻ cho nhà cung cấp hạ tầng đám mây và nhà cung cấp dịch vụ phần mềm. Người dùng đầu cuối (end user) chỉ là người dùng phần mềm với các lựa chọn cấu hình khác nhau được cung cấp bởi phần mềm nên không có nhiều vai trò trong an toàn hệ thống. Người dùng cuối chỉ biết tin vào nhà cung cấp phần mềm và các cam kết của nhà cung cấp về trách nhiệm bảo mật. Thông thường các cam kết này có thể là điều khoản trong hợp đồng thuê bao phần mềm, như là: an toàn thông tin và chất lượng dịch vụ. Chúng thường bao gồm: dung lượng dữ liệu, toàn vẹn dữ liệu, chính

sách về phân tán, sao lưu và phục hồi dữ liệu khi có sự cố, độ tin cậy, tính riêng tư và an toàn mạng cùng với các cam kết khác về chất lượng dịch vụ như dung lượng đường truyền, tính sẵn dùng 13.

Ở mức này, các phần mềm được cung cấp trên nền web (web-based application). Các web này thường được đặt ở máy chủ ảo trên đám mây, cho nên chúng phải được kiểm tra bằng cách quét các yếu điểm web nhờ vào một ứng dụng quét nào đó, ví dụ như các phần mềm có các chức năng như trong công bố 15. Các tường lửa có thể được dùng để ngăn chặn các tấn công vào điểm yếu đã biết của các phần mềm nền web. Những công việc này thuộc về nhà cung cấp phần mềm hoặc đám mây, người dùng cuối nhiều lắm là tham gia vào lựa chọn các cấu hình (option) khác nhau mà thôi. Tình hình này có thể dẫn đến những lộn xộn trong cấu hình an toàn chung của hệ thống do tính chất đa thuê bao, kéo theo những lỗ hổng trong an toàn hệ thống. Vì vậy, các nhà cung cấp phải có những chính sách chung bắt buộc và cách kiểm soát sao cho những cấu hình an toàn, bảo mật phải nhất quán, chặt chẽ và không có lỗ hổng.

4.2 Vấn đề quản lí an toàn hệ thống

Phần trên vừa trình bày cho thấy tính phức tạp trong "kỹ thuật" an toàn trên đám mây từ góc độ kiến trúc và dịch vụ của điện toán đám mây. Phần này xin đề cập đến một số khía cạnh về quản lí, vốn không thể tách rời với kỹ thuật nhằm đảm bảo cho sự áp dụng chính sách bảo mật đúng đắn, cộng tác và có trách nhiệm giữa các bên có liên quan trong điện toán đám mây. Nghiên cứu về quản lí an toàn trên đám mây là rất phức tạp vì nó liên quan đến số lượng lớn người có liên quan với các yêu cầu (requirement) khác nhau về an toàn. Việc quản lí an toàn liên quan tới việc xây dựng các yêu cầu về an toàn, đặc tả chính sách an toàn, cơ chế kiểm soát và các cấu hình khác nhau về an toàn tương ứng với các chính sách đặc thù. Việc quản lí này là động vì luôn phải đáp ứng các yêu cầu mới, các phản hồi từ môi trường và từ chính quá trình kiểm soát an toàn.

Điện toán đám mây cung cấp các dịch vụ trên cơ sở một hợp đồng trách nhiệm (SLA – Service Level Agreement), đây là pháp lý quan trọng trong các tranh chấp, bất đồng sau này. Hợp đồng thông thường bao gồm chất lượng dịch vụ, tính sẵn dùng, độ tin cậy và an toàn. Và như bao cam kết hợp đồng, nó có những điều khoản về trả phí dịch vụ xử phạt và bồi thường. Một đòi hỏi cao về an toàn thường dẫn đến một tiêu tốn nhiều nguồn lực và vì vậy mức giá dịch vụ sẽ cao lên tương ứng. Một mẫu (template) hướng dẫn về hợp đồng dịch vụ đám mây có thể tìm thấy tại 16. Độc giả quan tâm sâu phần này có thể xem thêm trong 17. Cần lưu ý rằng những đòi hỏi khắc khe về an toàn có thể ảnh hưởng đến hiệu năng chung của hệ thống (cồng kềnh hơn, chậm hơn,…) . Vì vậy, cần có sự cân bằng giữa yêu cầu an toàn, chi phí và hiệu năng của hệ thống. Nhiều khía cạnh khác liên quan đến quản lí an toàn cũng như các cơ quan quản lí an toàn đám mây có thể tìm thấy trong công bố 18.

5 KẾT LUẬN

Mặc dù điện toán đám mây đang được coi là một cuộc cách mạng Internet làm thay đổi cách ứng dụng công nghệ thông tin, nhưng việc chấp nhận nó vẫn còn nhiều vấn đề và e ngại chung quanh câu hỏi an toàn, bảo mật thông tin. Lợi ích của điện toán đám mây là rõ ràng và vô cùng hấp dẫn, nó làm giảm nhẹ chi phí đầu tư và gánh nặng bảo trì phần cứng, phần mềm, tuy nhiên từ kiến trúc, dịch vụ và các đặc điểm của điện toán đám mây cho thấy vẫn còn nhiều câu hỏi đặt ra cho vấn đề an toàn và bảo mật. Các vấn đề bảo mật ở cấp càng thấp thì vai trò và trách nhiệm của nhà cung cấp càng lớn, nhưng khách hàng có thể cảm thấy bất an vì họ không nắm rõ. Điều này có thể khắc phục bằng các hợp đồng (SLA) rõ ràng, chặt chẽ và tin cậy. Vấn đề an toàn có thể liên quan tới máy chủ ảo, bộ ảo hóa cũng như là kiến trúc hướng dịch vụ SOA.

Mặt khác, vấn đề an toàn trên điện toán đám mây không chỉ là trách nhiệm của nhà cung cấp dịch vụ mà còn là trách nhiệm của tất cả các bên có liên quan trong đám mây: nhà cung cấp, khách hàng, người dùng cuối. Vấn đề này có lẽ vẫn còn phải cần một thời gian nữa để có thể có giải pháp thỏa đáng làm tăng độ an toàn của đám mây, nhất là đám mây công cộng (public).

Điện toán đám mây còn rất mới và còn tiềm năng phát triển và ứng dụng, vấn đề an toàn của đám mây cần được nghiên cứu tiếp tục để ngày càng trở nên an toàn hơn. Mặt khác, sử dụng đám mây như thế nào cho có lợi, cân bằng giữa lợi ích và tính an toàn là sự tính toán của các nhà quản lí công ty, doanh nghiệp và sự tư vấn sáng suốt của các chuyên gia công nghệ thông tin.

TÀI LIỆU THAM KHẢO

1. Peter Mell, Timothy Grance, The NIST Definition of Cloud Computing, National Institute of Standards and Technology, Special Publication 800-145, September 2011.

2. http://en.wikipedia.org/wiki/Cloud_computing
3. Frank Gens, Robert P Mahowald and Richard L Villars, IDC Cloud Computing 2010.
4. http://www.idc.com/prodserv/idc_cloud.jsp
5. Chang, L, Ti ; Chin L; Chang, A.Y.; Chun J, C;(2010), Information security issue of enterprises adopting the application of cloud computing, IEEE 2010 Sixth International Conference on Networked Computing and Advanced Information Management (NCM), pp.645, 16-18 Aug. 2010.
6. Cloud Computing Use Case Discussion Group, "Cloud Computing Use Cases Version 3.0," 2010.
7. ENISA, "Cloud computing: benefits, risks and recommendations for information security," 2009,
8. http://www.enisa.europa.eu/act/rm/files/deliverables/cloud-computingrisk-assessment
9. Balachandra Reddy Kandukuri, Ramakrishna Paturi and Atanu Rakshit, "Cloud Security Issues," in Proceedings of the 2009 IEEE International Conference on Services Computing, 2009, pp. 517-520.
10. Kresimir Popovic , Zeljko Hocenski, "Cloud computing security issues and challenges," in The Third International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services, 2010, pp. 344-349.
11. Meiko Jensen, Jörg Schwenk, Nils Gruschka and Luigi Lo Iacono, "On Technical Security Issues in Cloud Computing," in IEEE ICCC, Bangalore 2009, pp. 109-116.
12. Bernd Grobauer, Tobias Walloschek and Elmar Stöcker, "Understanding Cloud-Computing Vulnerabilities," IEEE Security and Privacy, vol. 99, 2010.
13. S. Subashini, ,Kavitha, V., "A survey on security issues in service delivery models of cloud computing," Journal of Network and Computer Applications, Vol. 34, 2011.
14. Z. Wenjun, "Integrated Security Framework for Secure Web Services," in IITSI 2010 , pp. 178-183.
15. F. Elizabeth, , Vadim, Okun, "Web Application Scanners: Definitions and Functions," in HICSS 2007, pp. 280b-280b.
16. http://www.ibm.com/developerworks/cloud/library/cl-slastandards/
17. Wieder, P.; Butler, J.M.; Theilmann, W.; Yahyapour, R. (Eds.), Service Level Agreements for Cloud Computing, Springer, XXII, 2011.
18. Mohamed Al Morsy, John Grundy and Ingo Müller, An Analysis of The Cloud Computing Security Problem, Proceedings of APSEC 2010 Cloud Workshop, Sydney, Australia, 2010.
19. http://www.clouds360.com/iaas.php

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

ĐÁNH GIÁ KẾT QUẢ HỌC TẬP CỦA SINH VIÊN DỰA TRÊN TIẾP CẬN TRUNG BÌNH TRỌNG SỐ KẾT HỢP VỚI TRUNG BÌNH TRỌNG SỐ ĐƯỢC SẮP THỨ TỰ

Nguyễn Thị Thủy Chung[1] và Huỳnh Xuân Hiệp[1]
[1] *Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Evaluating student's study result based on weighted averaging combined with ordered weighted averaging approach

Từ khóa:
Phép toán OWA, phép toán WOWA, trung bình, độ đo, điểm thi

Keywords:
OWA (Ordered Weighted Averaging),W OWA (Weighted Ordered Weighted Averaging), Averaging , Ordered Weighted, measure, scores

ABSTRACT

This article focuses on evaluating students' study results based on weighted average combined with ordered weighted averaging by using WOWA operator (Weighted Ordered Weighted Averaging operator). The research also concentrates on the theory and application problems of the OWA operator (Ordered Weighted Averaging operator) and the extended WOWA operator aims to solving some practical issues based on the study results of the undergraduate students. The weights obtained in this research is determined based on the input information (ordered by statistics) for easy observation, and also for emphasizing the importance of considerably high (or low) values when collecting data of students. Together with the known average calculating method, this study also provides a new method to identify the average based on statistics, which is applied to evaluate the students' study results by combining weighted average (widely used) with the ordered weighted average. Early results of this study has proved the importance of applying integrated operators when collecting information from varied sources, criterions and purposes.

TÓM TẮT

Bài viết này nghiên cứu việc đánh giá kết quả học tập của sinh viên dựa trên tiếp cận trung bình trọng số kết hợp với trung bình trọng số được sắp thứ tự thông qua phép toán WOWA (Weighted Ordered Weighted Averaging operator). Nghiên cứu cũng tập trung vào các vấn đề lý thuyết và ứng dụng của phép toán tích hợp trung bình trọng số được sắp thứ tự OWA (Ordered Weighted Averaging operator) và phép toán mở rộng WOWA nhằm giải quyết một số vấn đề thực tiễn trên dữ liệu kết quả học tập sinh viên bậc đại học. Trọng số trong nghiên cứu này được xác định dựa trên thông tin đầu vào (sắp xếp theo thứ tự thống kê) để có thể dễ dàng quan sát, nhấn mạnh được mức độ quan trọng của các giá trị cao (hoặc thấp) cần quan tâm đánh giá trong quá trình tích hợp dữ liệu về sinh viên trong quá trình học tập bậc đại học. Cùng với các phương pháp tính trọng số đã biết, nghiên cứu cũng bổ sung một phương pháp mới dùng xác định trọng số dựa trên thống kê ứng dụng vào bài toán tính kết quả học tập của sinh viên kết hợp giữa trung bình trọng số (đang được áp dụng rộng rãi hiện nay) với trung bình trọng số được sắp thứ tự. Kết quả ban đầu của nghiên cứu đã khẳng định được tầm quan trọng của việc ứng dụng các phép toán tích hợp trong quá trình tích hợp thông tin từ nhiều nguồn khác nhau, đa tiêu chí, đa mục tiêu.

1 GIỚI THIỆU

Tích hợp là mối quan tâm cơ bản của tất cả các hệ thống kiến thức cơ sở 12357915. Tích hợp nhằm mục đích đồng thời sử dụng nhiều phần thông tin khác nhau (được cung cấp bởi nhiều nguồn) để cho ra một kết quả cụ thể. Phép toán tích hợp được thực hiện trên những đối tượng số học có chức năng biến đổi một tập những đối tượng này thành một số đại diện duy nhất. Có rất nhiều phép toán tích hợp khác nhau, mỗi một phép toán có những ứng dụng trong những lĩnh vực khác nhau tùy thuộc vào bài toán cụ thể. Các phép toán tích hợp mà chúng ta quen dùng và được ứng dụng rộng rãi là: trung bình số học (Arithmetic Mean) 15, trung bình trọng số (weighted mean)15, giá trị lớn nhất (Maximum)15, giá trị nhỏ nhất (Minimum)15. Một phép toán tích hợp đã được Yager đề nghị năm 1988 là trung bình trọng số được sắp thứ tự OWA (Ordered Weighted Averaging operator) 7, OWA chính là phép toán tổng quát khái quát các phép toán như trung bình số học, giá trị lớn nhất, giá trị nhỏ nhất, trung vị và thống kế thứ tự thứ k (k-order statistics).

Thông thường, phép toán tích hợp được sử dụng để tính điểm trung bình ở các trường đại học là trung bình trọng số, trong đó trọng số chính là số tín chỉ của từng môn học. Do đó điểm trung bình phụ thuộc vào số tín chỉ, những môn nào có số tín chỉ càng cao thì trọng số càng lớn. Tuy nhiên, tính điểm trung bình theo phép toán trung bình trọng số như hiện nay không làm nổi bật được mức độ quan trọng của giá trị điểm thi. Ví dụ: một sinh viên học ba môn có điểm lần lượt là 10, 8, 4 khi tính trung bình làm sao thể hiện được mức độ quan trọng của điểm cao hơn, giả sử trong ba giá trị điểm trên thì điểm cao hơn phải có hệ số lớn hơn. Phép toán tích hợp trung bình trọng số được sắp thứ tự (trong nghiên cứu này sẽ gọi là phép toán tích hợp OWA) và các phép toán mở rộng WOWA của phép toán tích hợp OWA sẽ giúp ta giải quyết các vấn đề trên. Phép toán tích hợp OWA sẽ kết hợp thông tin trong đó cho phép giá trị của trọng số liên kết với thứ tự của vị trí chứ không liên kết với thông tin nguồn. Bằng phép toán OWA, việc đánh giá có thể giảm bớt độ quan trọng của những giá trị ở hai đầu mút và tăng độ quan trọng của những phần tử ở trung tâm, hoặc tăng độ quan trọng của những giá trị ở đầu mút nào đó tùy theo mục đích đánh giá dữ liệu. Dựa trên kết quả đánh giá này ta có thể cho ra các mô hình quyết định khác nhau. Phép toán WOWA cho phép kết hợp phép toán trung bình trọng số và phép toán OWA.

Phép toán tích hợp OWA là một công cụ hữu ích nhằm tích hợp các thuộc tính của đối tượng theo các tiêu chí khác nhau. Phép toán này đã được sử dụng trong nhiều dạng bài toán và đã thu được những kết quả tốt. Việc áp dụng phép toán tích hợp OWA để mô hình hóa quyết định cũng đã được ứng dụng trong nhiều lĩnh vực. Năm 2006, Chang đề nghị một mô hình fuzzy OWA động liên quan đến vấn đề ra quyết định đa tiêu chí (Multiple Criteria Decision Making – MCDM) 1. Mô hình này giúp người sử dụng có thể giải quyết vấn đề quyết định đa tiêu chí dựa vào các thông tin mờ hoặc thông tin không hoàn tất. Phép toán OWA cũng có thể được áp dụng hiệu quả trong lĩnh vực quản trị nguồn nhân lực, năm 2008, L.Canós và V. Liern đã phát triển một hệ thống hỗ trợ ra quyết định linh động để giúp những nhà quản lý trong việc thực hiện chức năng ra quyết định chọn lựa nhân sự của họ. Hệ thống này mô phỏng đánh giá của chuyên gia trên cơ sở dùng phép toán tích hợp OWA để gán những trọng số khác nhau vào những tiêu chí chọn lựa khác nhau, từ đó sử dụng mô hình tích hợp dựa trên các phân tích hiệu quả để sắp xếp các ứng viên theo thứ tự 5. Gần đây, năm 2012, José M. Merigó và Anna M. Gil-Lafuente cũng đã đề xuất một hướng nghiên cứu mới trong việc dùng phép toán OWA kết hợp với trung bình trọng lượng trong cùng một công thức để hỗ trợ cho việc ra quyết định trong lĩnh vực khá nhạy cảm như kinh doanh và kinh tế, thay vì chỉ dùng các toán tử tích hợp thông thường như khoảng cách Hamming trước đây 4. Ngoài ra còn rất nhiều vấn đề trong các lĩnh vực khác được giải quyết bằng phép toán tích hợp OWA và các phép toán mở rộng của OWA, trong bài viết này sẽ đề cập đến phép toán mở rộng của OWA là phép toán WOWA.

Bài viết được tổ chức thành 6 phần. Phần một giới thiệu chung về phép toán tích hợp OWA để mô hình hóa trong một số lĩnh vực. Phần hai trình bày cụ thể về phép toán OWA, định nghĩa và các thuộc tính. Phần ba trình bày các độ đo và các phương pháp xác định trọng số cho từng môn học của sinh viên trong từng học kỳ theo phương pháp thống kê và phương pháp lượng hóa. Phần bốn giới thiệu phép toán mở rộng WOWA. Phần năm tập trung trình bày cách tiếp cận dữ liệu đào tạo, kết quả thực nghiệm của phép toán WOWA trên dữ liệu sinh viên, phân tích các kết quả có được trên cơ sở sử dụng công cụ minh họa kết quả học tập tính bằng phép toán tích hợp trung bình trọng số được sắp thứ tự. Tóm tắt một số kết quả quan trọng và hướng phát triển được nêu ra ở phần cuối cùng.

2 PHÉP TOÁN TÍCH HỢP TRUNG BÌNH TRỌNG SỐ ĐƯỢC SẮP THỨ TỰ (OWA)

2.1 Giới thiệu phép toán OWA

Việc tích hợp một tập những đối tượng để hình thành một đối tượng mới có tầm quan trọng đáng kể trong nhiều lĩnh vực. Một nhân tố chính trong việc xác định cấu trúc của hàm tích hợp là sự tương quan giữa các đối tượng thỏa mãn các yêu cầu đặt ra. Phép toán tích hợp OWA (Ordered Weighted Averaging) được giới thiệu lần đầu bởi Yager nhằm cung cấp một phương pháp tích hợp các đối tượng thành một đối tượng mới 7. Theo phương pháp tích hợp này thì phép toán trung bình trọng số là một trường hợp đặc biệt của phép toán tích hợp OWA. Nhưng phép toán tích hợp OWA khác với trung bình trọng số ở chỗ trọng số không liên kết trực tiếp với giá trị thực mà liên kết với thứ tự của vị trí các đối tượng cần tích hợp. Ví dụ ta cần tích hợp điểm thi của sinh viên bằng phép toán tích hợp OWA thì trọng số không liên kết trực tiếp với điểm thi của từng môn cụ thể mà liên kết với vị trí của điểm thi sau khi đã sắp xếp.

2.2 Định nghĩa phép toán OWA

Một phép toán tích hợp trung bình trọng số được sắp thứ tự (OWA)715 *n* chiều là một ánh xạ $F: R^n \to R$ với véctơ trọng số $W = (w_1, w_2, \dots, w_n)$ thỏa các điều kiện sau:

$\mathbf{w_i \in [0,1]}$ với $\mathbf{1 \le i \le n}$

$$\sum_{i=1}^{n} \mathbf{w_i} = \mathbf{1}$$

$$F(a_1, a_2, \dots, a_n) = \sum_{i=1}^{n} w_i b_i$$

Với b_i là phần tử có giá trị lớn thứ *i* được chọn từ tập $a_1, a_2, \dots, a_n$, $b_i \in \{a_1, a_2, \dots, a_n\}$ hay $(b_1, b_2, \dots, b_n)$ là một hoán vị của $(a_1, a_2, \dots, a_n)$.

Chúng ta thấy phép toán trung bình trọng số được sắp thứ tự (OWA)7 khác với trung bình trọng số (WM)15 ở chỗ trọng số không liên kết trực tiếp với giá trị thực a_i mà liên kết với giá trị b_i theo thứ tự của vị trí. B được gọi là vectơ tham số đã được sắp thứ tự nếu $\forall b_i, b_i \ge b_j, i < j$.

Với F là phép toán tích hợp OWA với vectơ trọng số W và một bộ tham số $(a_1, a_2, \dots, a_n)$, ta có thể liên kết với bộ tham số này một vectơ B đầu vào đã sắp thứ tự, như vậy B là vectơ bao gồm các tham số của F đặt theo thứ tự giảm dần.

Ký hiệu $F(a_1, a_2, \dots, a_n) = WB$. Ta có thể biểu thị $F(a_1, a_2, \dots, a_n)$ như $F(B)$, với B là sự kết hợp của vectơ tham số đã sắp thứ tự. Điều quan trọng cần nhấn mạnh là các trọng số được liên kết với một vị trí cụ thể có thứ tự chứ không liên kết với một phần tử đặc biệt, đó là trọng số w_i liên kết với phần tử lớn nhất thứ i.

Tính tổng quát của phép toán tích hợp OWA là nó có thể thực hiện phép toán tích hợp để cho ra các kết quả khác nhau bằng cách chọn lựa trọng số khác nhau. Đặc biệt, bằng cách chọn lựa các trọng số thích hợp trong vectơ W nó có thể nhấn mạnh các tham số khác nhau trên cơ sở vị trí của chúng trong sắp thứ tự. Nếu ta đặt hầu hết các trọng số gần đỉnh của W ta có thể nhấn mạnh các tham số có giá trị cao hơn, trong khi đó nếu ta đặt các trọng số gần đáy của W sẽ nhấn mạnh các tham số có giá trị thấp hơn trong phép tích hợp.

3 CÁCH XÁC ĐỊNH TRỌNG SỐ

Ta biết rằng phép toán OWA phụ thuộc vào vectơ trọng số. Có nhiều phương pháp đã được đề nghị để xác định trọng số trong phép toán OWA.

3.1 Xác định trọng số bằng phép lượng hóa tích hợp

Xác định trọng số bằng phép lượng hóa tích hợp 7 là một trong những phương pháp cho phép xác định trọng số mà kết quả của nó thể hiện được mức độ quan trọng của giá trị đã được sắp xếp. Phương pháp này được thực hiện như sau:

$$w_i = \left(\frac{i}{n}\right)^{\alpha} - \left(\frac{i-1}{n}\right)^{\alpha}, i = 1..n$$

Trong đó hệ số α được xác định tùy theo bài toán cụ thể (xem ví dụ cách xác định hệ số α ở (101415)

Do đó:

$$\mathbf{orness(W) < 0.5}, \quad \alpha > 1$$

$$orness(W) = 0.5, \quad \alpha = 1$$

$$\mathbf{orness(W) > 0.5}, \quad \alpha < 1$$

3.2 Xác định trọng số bằng phương pháp thống kê: dựa vào giá trị của các tham số để tạo ra trọng số thỏa các điều kiện của phép toán OWA

Phương pháp thống kê cho phép xác định trọng số dựa trên tập dữ liệu gồm nhiều bộ giá trị. Phương pháp này tạo ra các trọng số w_i tương ứng với thành phần thứ i của mỗi bộ dữ liệu. Các giá trị trọng số vẫn đảm bảo các điều kiện của phép toán OWA là trọng số phải thể hiện được ý nghĩa sắp thứ tự của các giá trị trong từng bộ dữ liệu (có nghĩa là trọng số thứ i gắn liền với vị trí thứ i), và tổng trọng số phải bằng 1.

Xét tập dữ liệu đầu vào gồm m bộ tham số. Mỗi bộ gồm có n thành phần với $a_{j,i}$ là thành phần thứ i của bộ thứ j .

Ký hiệu $\mathbf{tongcot}_i$ là tổng giá trị lớn thứ i của tất cả các bộ tham số $a_{j,i}$ với i = 1..n và j = 1..n.

Ta có:
$$tongcot_i = \sum_{j=1}^{m} a_{j,i}$$

Ký hiệu $tong$ là tổng tất cả các giá trị tham số đầu vào. Ta có:
$$tong = \sum_{j=1}^{m}\sum_{i=1}^{n} a_{j,i}$$

Trọng số của thành phần thứ i được xác định như sau:
$$w_i = \frac{tongcot_i}{tong}$$

Tổng quát ta có công thức tính trọng số như sau:
$$w_i = \frac{\sum_{j=1}^{m} a_{j,i}}{\sum_{j}^{m}\sum_{k}^{n} a_{jk}}$$

Theo phương pháp thống kê trọng số sẽ giảm dần, có nghĩa là: $w_i \geq w_{i+1}$.

Tính trọng số theo phương pháp thống kê thể hiện được tỉ lệ của giá trị lớn thứ i trên tổng tất cả các giá trị, do đó nó thể hiện được mức độ quan trọng lớn nhỏ của từng thành phần. Nói cách khác, xét trên từng bộ dữ liệu nếu thành phần thứ i có giá trị lớn hơn thành phần thứ j, thì mức độ quan trọng của thành phần thứ i sẽ lớn hơn thành phần thứ j. Phương pháp thống kê tạo ra một bộ trọng số thỏa các ràng buộc về trọng số của phép toán tích hợp sắp thứ tự OWA là trọng số thể hiện được tính chất sắp thứ tự của dữ liệu và tổng trọng số luôn bằng 1.

4 PHÉP TOÁN WEIGHTED ORDERED WEIGHTED AVERAGING OPERATOR (WOWA)

Mặc dù cả hai phép toán trung bình trọng số và phép toán tích hợp OWA dùng để kết hợp những tham số theo một tập các trọng số. Tuy nhiên ý nghĩa của những trọng số này thì khác nhau trong cả hai phép toán. Trung bình trọng số tính giá trị của các tham số mà giá trị của các tham số có xem xét đến tính tin cậy của thông tin nguồn. Nghĩa là, mỗi giá trị được gán một trọng số dựa vào mức độ tin cậy của nguồn cung cấp nó. Trọng số càng lớn thì mức độ ảnh hưởng của giá trị tương ứng với kết quả sau cùng càng nhiều [8]. Còn phép toán tích hợp OWA tích hợp thông tin mà nó cho phép giá trị của trọng số liên kết với thứ tự của vị trí (vị trí của các tham số sau khi đã sắp xếp). Bằng cách này, một hệ thống có thể giảm bớt độ quan trọng của những giá trị ở hai đầu mút và tăng độ quan trọng của những phần tử ở trung tâm.

Như vậy, khi xem xét riêng lẻ hai phép toán trung bình trọng số và phép toán tích hợp OWA [9], mỗi phép toán chỉ thể hiện được ưu điểm riêng của trọng số trong từng trường hợp. Ví dụ: Để tính điểm trung bình các môn học trong từng học kỳ thì điểm thi có thể được kết hợp bằng phép toán trung bình trọng số (với trọng số được xác định bằng số tín chỉ chính là thông tin nguồn mà ta đưa vào). Tuy nhiên, với tính độc lập của số tín chỉ của từng môn học, nó sẽ cung cấp những giá trị mà độ quan trọng không bằng nhau. Trong trường hợp ta muốn những môn có điểm thi cao có độ quan trọng lớn hơn (hoặc nhỏ hơn) nhưng vẫn tôn trọng mức độ quan trọng của số tín chỉ của từng môn học thì phép toán "Weighted OWA" (WOWA) sẽ giúp ta giải quyết được vấn đề trên. Phép toán WOWA cho phép người sử dụng xem xét cả hai khía cạnh của những tập các trọng số đó là: xác định trọng số chính xác với thông tin nguồn như trung bình trọng số và giá trị liên quan đến vị trí như phép toán tích hợp OWA.

4.1 Định nghĩa13

Cho P và W là các vectơ trọng số n chiều: $P = (p_1 \ldots p_n)$ và $W = (w_1, \ldots w_n)$ thỏa:

i) $p_i \in [0,1]$ và $\sum_{i=1}^{n} p_i = 1$

ii) $w_i \in [0,1]$ và $\sum_{i=1}^{n} w_i = 1$

Một ánh xạ F_{wowa}: $R^n \rightarrow R$ được gọi là phép toán "Weighted Ordered Weighted Averaging - WOWA" n chiều nếu:

$$F_{wowa} = \sum_{i=1}^{n} \omega_i b_j$$

với b_j là phần tử lớn nhất thứ j của $\{a_1, a_2, \ldots, a_n\}$ và trọng số $\omega_i = W^*(\sum_{j \le i} p_{\sigma(j)}) - W^*(\sum_{j<i} p_{\sigma(j)})$, trong đó $p_{\sigma(j)}$ là phần tử lớn nhất thứ j của $\{p_1, p_2, \ldots\}$

Với W* là hàm đơn điệu tăng dần được xây dựng bằng cách nội suy những điểm: $(\frac{i}{n}, \sum_{j \le i} w_j)_{i=1,\ldots,n} \cup (0,0)$. W* yêu cầu phải nội suy tuyến tính khi những điểm có thể được nội suy theo cách này.

Trong định nghĩa trên vectơ trọng số P tương ứng với vectơ trọng số trong phép toán trung bình số học, W tương ứng với vectơ trọng số trong phép toán OWA. Còn ω_i là một vectơ trọng số mới dùng để xem xét sự tương tác giữa p và q.

– Khi $P = (\frac{1}{n}, \ldots, \frac{1}{n})$ phép toán tích hợp WOWA là phép toán tích hợp OWA:

$WOWA_{p,w}(a_1,..,a_n) = OWA_w(a_1,...,a_n)$.

– Khi $W = (\frac{1}{n}, \ldots, \frac{1}{n})$ phép toán WOWA là phép toán trung bình trọng số:

$WOWA_{p,w}(a_1,...,a_n) = WM_p(a_1,...,a_n)$.

– Khi $P=W = (\frac{1}{n}, \ldots, \frac{1}{n})$ thì phép toán WOWA chính là phép toán trung bình số học:

$WOWA_{p,w}(a_1,\ldots,a_n) = AM(a_1,...,a_n)$.

4.2 Phương pháp tính trọng số W*

Phép toán tích hợp WOWA định nghĩa vectơ trọng số ω dựa vào sự khác nhau giữa các cặp điểm trong hàm W*. Các điểm được chọn bằng cách sử dụng vectơ P và xây dựng hàm bằng cách sử dụng véctơ W 15.

Phép toán WOWA là khái quát hóa của phép toán OWA, các trọng số ω_i bằng với trọng số w_i khi tất cả các nguồn thông tin có độ quan trọng như nhau, có nghĩa là khi tất cả các trọng số p_i bằng nhau ($p_i = \frac{1}{n}$).

Gọi p^0 là vectơ trọng số mà độ quan trọng của p_i như nhau:

$$p^0 = (p_1^0, \ldots, p_n^0) = (\frac{1}{n}, \ldots, \frac{1}{n})$$

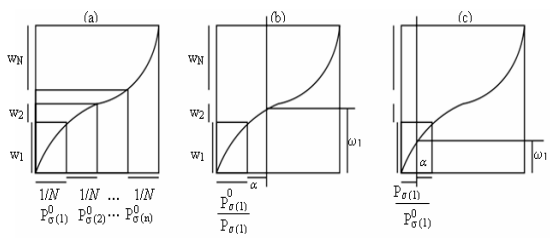
Chú ý: khi $p = p^0$ sẽ không có sự tương tác giữa các trọng số, vì vậy $p^0_{\sigma(1)} = \frac{1}{n}$ và w_1 tham chiếu tương ứng đến trọng số của thông tin nguồn đó là giá trị lớn nhất của $\{a_1,\ldots,a_n\}$ và trọng số cho giá trị này. Tương tự, $p^0_{\sigma(2)} = \frac{1}{n}$ và w_2 tham chiếu tương ứng đến giá trị lớn thất thứ 2 của $\{a_1,\ldots,a_n\}$. Một cách tổng quát: w_i và $p_{\sigma(i)}$ là trọng số tham chiếu đến giá trị $a_{\sigma(i)}$ với σ là một hàm hoán vị các giá trị i sao cho $a_{\sigma(i)}$ là phần tử lớn nhất thứ i của $\{a_1, a_2, \ldots, a_n\}$.

Khảo sát quan hệ giữa W và P trên đồ thị. Xét tập hợp các điểm:

$$\{(\sum_{j \le i} p^0_{\sigma(j)}, \sum_{j \le i} w_j)\}_{i=1,n} = \{(\sum_{j \le i} p^0_{\sigma(j)} = \sum_{j \le i} \frac{1}{n} = \frac{i}{n}, \sum_{j \le i} w_i)\}_{i=1,n}$$

Khoảng cách giữa 2 điểm liên tiếp trên trục y là w_i, khoảng cách giữa 2 điểm liên tiếp trên trục x là $p^0_{\sigma(i)}$. Do $\sum w_i = 1$ và $\sum p_i^0 = 1$ nên các điểm này nằm trong khoảng đơn vị, và do $w_i \ge 0$, $p_i^0 \ge 0$ nên nó có hình dạng của hàm đơn điệu.

W* được định nghĩa như một hàm đơn điệu và nội suy các điểm và là hàm nội suy theo đường thẳng. Hình 1(a) biểu diễn những điểm trên hàm W*15.



Hình 1: Xây dựng trọng số ω của WOWA

(a) Xây dựng hàm w*

(b) Tính ω_1 khi $p_{\sigma(1)} > p^0_{\sigma(1)}$

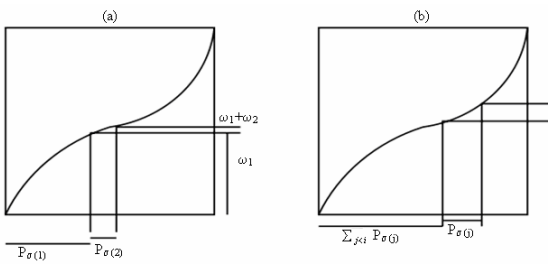
(c) Tính ω_1 khi $p_{\sigma(1)} < p^0_{\sigma(1)}$

Việc chọn trọng số ω từ đường cong thì giống như việc mở rộng hoặc thu hẹp đoạn p_i^0 trên trục x, sự thay đổi này tạo ra các điểm trượt trên đường cong đồ thị w* và ω chính là khoảng thay đổi này trên trục y.

Xét trường hợp phần tử lớn nhất $a_{\sigma(i)}$ và trọng số w_1, $p_{\sigma(i)}$. Trong trường hợp này, nếu $p_{\sigma(1)} > p^0_{\sigma(1)} = \frac{1}{n}$ thì $\omega_1 > w_1$. Chú ý: nếu tăng $p_{\sigma(1)}$ đến $p^0_{\sigma(1)}$ (nghĩa là: $p_{\sigma(1)} = p^0_{\sigma(1)} + \alpha$ với α > 0) thì $w^*(p_{\sigma(1)}) > w^*(p^0_{\sigma(1)})$. Giá trị $w^*(p_{\sigma(1)})$ tương đương với trọng số $\omega_1 = w^*(p_{\sigma(1)}) = w^*(p^0_{\sigma(1)} + \alpha)$ thể hiện trong Hình 1(c).

Tiếp tục xem xét phần tử lớn thứ hai $a_{\sigma(2)}$ với trọng số p tương ứng là $p_{\sigma(2)}$. Trong trường hợp này giá trị cộng dồn $p^0_{\sigma(1)} + p^0_{\sigma(2)}$ tương ứng với $p_{\sigma(1)} + p_{\sigma(2)}$. Tính $w^*(p_{\sigma(1)} + p_{\sigma(2)})$, hàm này tương ứng với giá trị $\omega_1+\omega_2$ trong Hình 2(a), và vì vậy $w_2 = w^*(p_{\sigma(1)} + p_{\sigma(2)}) - w^*(p_{\sigma(1)})$. Chú ý: giá trị ω_2 không chỉ phụ thuộc vào $p_{\sigma(2)}$ mà còn phụ thuộc vào $p_{\sigma(1)}$.

Tương tự ta tính các giá trị ω_i khác. Hình 2(b) chỉ ra cách tính ω_i 15. Như ω_2, ta thấy ω_i không chỉ phụ thuộc vào $p_{\sigma(i)}$ mà còn phụ thuộc vào $p_{\sigma(j)}$ và w_j với $j \le i$



Hình2: Cách tính ω_i

5 THỰC NGHIỆM

Dữ liệu về điểm thi của sinh viên được lưu trữ theo từng học kỳ. Trong bài viết này, chúng tôi lấy dữ liệu của 1 lớp gồm có 25 sinh viên, mỗi sinh viên học 8 môn (xem phụ lục để biết cấu trúc bảng điểm thi gồm có 4 cột: mã sinh viên, mã môn, số tín chỉ, điểm thi) để đánh giá kết quả học tập của sinh viên. Mục đích của thực nghiệm là so sánh kết quả học tập được đánh giá theo trung bình trọng số và trung bình trọng số kết hợp với trung bình trọng số được sắp thứ tự (WOWA). Ngoài trong thực nghiệm cũng chỉ ra điểm nổi bật của trọng số được tính theo phương pháp thống kê đã được chỉ ra trong nghiên cứu.

5.1 Đánh giá kết quả học tập bằng phép toán WOWA với trọng số được xác định bằng phương pháp lượng hóa:

Quan sát kết quả học tập được đánh giá bằng phép toán WOWA với những hệ số α khác nhau trong Bảng 1:

Bảng 1: Tính trung bình điểm thi bằng phép toán WOWA

MASV	TBTS	TRUNG BÌNH WOWA VỚI α KHÁC NHAU		
		1.5	1.0	0.3
1050538	6,05	6,05	6,05	5,15
1050541	6,25	6,40	6,25	5,10
1050551	8,13	8,13	8,13	7,83
1050557	7,92	7,85	7,92	7,84
1050578	7,75	7,83	7,75	7,06
1050589	7,69	7,71	7,69	7,79
1050590	7,78	7,55	7,78	8,46
1050597	7,22	7,16	7,22	7,27
1050603	7,56	8,24	7,56	5,95
1050607	6,16	5,89	6,16	7,97
1050654	7,53	7,51	7,53	8,05
1050658	7,61	7,41	7,61	8,03
1050662	8,34	8,37	8,34	8,59
1050666	6,62	6,70	6,62	7,95
1050678	2,88	3,07	2,88	1,57

Các hệ số α = 1,5, α = 1,0 hay α = 0,3 chỉ là những giá trị được chọn ngẫu nhiên lớn hơn, nhỏ hơn hay bằng 1, để so sánh kết quả của trung bình WOWA với trung bình trọng số. Kết quả ở Bảng 3 cho thấy khi hệ số α = 1 thì trung bình WOWA sẽ có giá trị bằng với trung bình trọng số (trọng số ở đây chính là số tín chỉ). Trong trường hợp này phép toán WOWA chính là sự kết hợp của phép toán trung bình trọng số với phép toán OWA, mà trong phép toán OWA khi hệ số α = 1 thì trọng số của từng môn thi sẽ bằng nhau (không quan tâm đến giá trị điểm thi cao hay thấp). Khi α <>1 ta thấy trung bình WOWA của từng sinh viên có thể nhỏ

hơn hoặc lớn hơn trung bình trọng số. Do trong phép toán OWA, khi α lớn hơn 1 thì những điểm có giá trị càng nhỏ sẽ có trọng số càng lớn. Ngược lại, khi α nhỏ hơn 1 thì những điểm có giá trị lớn hơn sẽ có trọng số lớn hơn. Phép toán WOWA là sự kết hợp của phép toán trung bình trọng số và phép toán OWA, nên trọng số của từng môn học phụ thuộc cả vào số tín chỉ và cả hệ số α (ở đây hệ số α thay đổi cho phép đánh giá "sức nặng" của giá trị điểm thi).

Xét cụ thể điểm thi của 2 sinh viên trong Bảng 2 như sau:

Bảng 2: Bảng điểm thi của 2 sinh viên

MASV	MÃ MÔN	TC	ĐIỂM
1050538	CT126	1	7,5
1050538	CT106	2	3,5
1050538	CT118	4	5
1050538	CT109	2	5,5
1050538	CT110	2	6
1050538	CT113	3	8,5
1050538	CT111	3	6
1050538	CT335	2	7
1050551	CT118	3	7,5
1050551	CT314	4	10
1050551	CT109	2	9
1050551	CT335	2	8
1050551	CT301	2	7
1050551	CT113	3	7,5
1050551	CT111	2	6
1050551	CT310	1	9,5

Với α khác nhau thì trọng số OWA của sinh viên "1050538" được xác định như Bảng 3:

Bảng 3: Trọng số OWA của sinh viên 1050538

Môn	Tín chỉ	Điểm	Trọng số OWA		
			α = 1,5	α = 1,0	α = 0,3
CT126	2	3,5	0,18	0,13	0,04
CT106	4	5	0,17	0,13	0,04
CT118	2	5,5	0,16	0,13	0,05
CT109	3	6	0,12	0,13	0,07
CT110	2	6	0,14	0,13	0,06
CT113	2	7	0,10	0,13	0,09
CT111	1	7,5	0,08	0,13	0,12
CT335	3	8,5	0,04	0,13	0,54

Qua bảng này ta thấy khi α > 1 thì điểm thi càng nhỏ sẽ có trọng số càng lớn. Ngược lại, khi α < 1, điểm thi lớn sẽ có trọng số lớn.

Trọng số OWA của sinh viên "1050551" được xác định như Bảng 4:

Bảng 4: Trọng số OWA của sinh viên 1050551

Môn	Tín chỉ	Điểm	Trọng số OWA		
			α = 1,5	α = 1,0	α = 0,3
CT118	2	6	0,18	0,13	0,04
CT314	2	7	0,17	0,13	0,04
CT109	3	7,5	0,14	0,13	0,06
CT335	3	7,5	0,16	0,13	0,05
CT301	2	8	0,12	0,13	0,07
CT113	2	9	0,10	0,13	0,09
CT111	1	9,5	0,08	0,13	0,12
CT310	4	10	0,04	0,13	0,54

Bảng 3 và Bảng 4 cho thấy 2 sinh viên cùng học môn CT335 nhưng trọng số OWA sẽ khác nhau do trọng số phụ thuộc vào thứ tự của điểm thi của từng sinh viên chứ không phụ thuộc vào số tín chỉ của môn học.

Trọng số WOWA của sinh viên "1050538" theo phương pháp lượng hóa được xác định như Bảng 5:

Bảng 5: Trọng số WOWA của sinh viên 1050538

Môn	Tín chỉ	Điểm	Trọng số WOWA		
			α = 1,5	α = 1,0	α = 0,3
CT126	2	3,5	0,12	0,16	0,14
CT106	4	5	0,27	0,21	0,15
CT118	2	5,5	0,04	0,11	0,45
CT109	3	6	0,11	0,11	0,06
CT110	2	6	0,13	0,11	0,04
CT113	2	7	0,10	0,11	0,06
CT111	1	7,5	0,18	0,16	0,07
CT335	3	8,5	0,06	0,05	0,02

Trọng số WOWA của sinh viên "1050551" được xác định như Bảng 6:

Bảng 6: Trọng số WOWA của sinh viên "1050551"

Môn	Tín chỉ	Điểm	Trọng số WOWA		
			α = 1,5	α = 1,0	α = 0,3
CT118	2	6	0,15	0,11	0,03
CT314	2	7	0,05	0,05	0,03
CT335	3	7,5	0,24	0,21	0,09
CT109	3	7,5	0,18	0,16	0,09
CT301	2	8	0,08	0,11	0,09
CT113	2	9	0,09	0,11	0,07
CT111	1	9,5	0,14	0,11	0,04
CT310	4	10	0,07	0,16	0,55

Bảng 5 và Bảng 6 cho thấy trọng số WOWA không chỉ phụ thuộc vào thứ tự của điểm thi của từng sinh viên mà còn phụ thuộc vào số tín chỉ của môn học.

Qua kết quả trên cho thấy, nếu tính điểm trung bình theo trung bình trọng số (đang được áp dụng ở Trường Đại học Cần Thơ) thì điểm trung bình được tính với trọng số là số tín chỉ, môn học có số tín chỉ càng lớn thì trọng số càng lớn và trọng số này là giống nhau đối với tất cả các sinh viên cùng học 1 môn. Bằng phép toán WOWA, vừa tôn trọng giá trị của số tín chỉ vừa thể hiện được sự ảnh hưởng của giá điểm cao hay thấp bằng cách thay đổi giá trị α để tăng trọng số cho điểm cao hay điểm thấp.

5.2 Đánh giá kết quả học tập bằng phép toán WOWA với trọng số được xác định bằng phương pháp thống kê:

Tính điểm trung bình bằng phép toán WOWA với trọng số xác định bằng phương pháp lượng hóa như trên vừa tôn trọng giá trị của số tín chỉ vừa thể hiện được sự ảnh hưởng của giá điểm cao hay thấp. Tuy nhiên phương pháp này chưa thấy được ảnh hưởng điểm thi của tất cả các sinh viên đối với từng sinh viên. Xét kết quả điểm trung bình được tính bằng phép toán WOWA với trọng số được xác định bằng phương pháp thống kê được đưa ra trong nghiên cứu này ta sẽ thấy điểm trung bình sẽ thay đổi tùy thuộc vào điểm thi của cả lớp.

Xét lớp có 10 sinh viên (lấy 10 sinh viên đầu trong bảng phụ lục), ta có điểm trung bình WOWA với trọng số xác định bằng phương pháp thống kê như Bảng 7:

Bảng 7: Trung bình WOWA tính trên 10 sinh viên theo phương pháp thống kê

MASV	TBTS	TB WOWA
1050538	6,05	6,05
1050541	6,25	6,27
1050551	8,13	8,02
1050557	7,92	8,03
1050578	7,75	7,84
1050589	7,69	7,67
1050590	7,78	7,82
1050597	7,22	7,18
1050603	7,56	7,30
1050607	6,16	6,42

Xét lớp có 25 sinh viên như bảng phụ lục, ta có kết quả như Bảng 8:

Bảng 8: Trung bình WOWA tính trên 25 sinh viên theo phương pháp thống kê

MASV	TBTS	TB WOWA
1050538	6,05	6,03
1050541	6,25	6,24
1050551	8,13	8,04
1050557	7,92	8,01
1050578	7,75	7,81
1050589	7,69	7,68
1050590	7,78	7,85
1050597	7,22	7,19
1050603	7,56	7,32
1050607	6,16	6,40
1050654	7,53	7,61
1050658	7,61	7,64
1050662	8,34	8,26
1050666	6,62	6,72
1050678	2,88	2,80
1050701	6,03	5,80
1050723	3,97	3,95
1050725	7,78	7,74
1050730	7,11	7,13
1050731	9,06	8,97
1050739	1,79	1,83
1050747	6,09	5,87
1050769	5,88	5,89
1050782	2,81	2,24

Quan sát Bảng 7 và Bảng 8 ta thấy, khi tính điểm trung bình bằng phép toán tích hợp WOWA (với trọng số xác định bằng phương pháp thống kê) thì điểm trung bình của một sinh viên sẽ thay đổi

khi tổng số sinh viên thay đổi. Nói cách khác điểm trung bình của sinh viên cũng bị ảnh hưởng bởi điểm thi của tất cả các sinh viên.

6 KẾT LUẬN

Qua kết quả sử dụng phép toán WOWA (phép toán trung bình trọng số kết hợp với phép toán tích hợp với trung bình trọng số sắp thứ tự) để tính kết quả học tập của sinh viên sẽ cho chúng ta thấy điểm trung bình không chỉ hoàn toàn phụ thuộc vào trọng số là số tín chỉ của từng môn học mà còn phụ thuộc vào thứ tự điểm thi của từng sinh viên. Ngoài ra việc xác định trọng số bằng phương pháp thống kê cho thấy tầm quan trọng của điểm thi cả lớp sẽ ảnh hưởng đến kết quả học tập của từng sinh viên.

Theo kết quả nghiên cứu trên, chúng tôi thấy việc sử dụng phép toán tích hợp WOWA với trọng số xác định bằng phương pháp thống kê rất phù hợp để đánh giá kết quả học tập của sinh viên được đào tạo theo qui chế tín chỉ và phù hợp với cách tính điểm theo thang điểm chữ.

TÀI LIỆU THAM KHẢO

1. Nguyễn Hoàng Phương, Nadipuram R.Prasad, Lê Minh Phong, *Nhập môn trí tuệ tính toán*, Nhà xuất bản Khoa học & Kỹ thuật, Hà Nội, p.71-90, 2002.
2. J.R.Chang, T.H.Ho, C.H.Cheng, A.P.Chen, "Dynamic fuzzy OWA model for group multiple criteria decision making", *Soft Computing,* p.543-554, 2006.
3. J.W.Wang, J.R Chang, CH Cheng (2006), "Flexible fuzzy OWA querying method for hemodialysis database", p.1031-1042.
4. José M. Merigó, Anna M. Gil-Lafuente (2012), "Decision-making techniques with similarity measures and OWA operators", SORT 36 January-June 2012, p. 81-102.
5. L.Canós, V. Liern (2008), "Soft computing-based aggregation methods for human resource management", European Journal of Operation Research 189, p.669-681, 2008.
6. Rehan Sadiq , S.T. (2007) "Probability density functions based weights for ordered weighted averaging (OWA) operators: An example of water quality indices" v.182(3): p. 1350-1368.
7. R.R.Yager, "On ordered weighted Averaging Aggregation Operators in Multicriteria Decisionmaking ". v.18(1), p. 183 – 190, 1988.
8. R.R.Yager, D.Filev (1999), "Induced Ordered Weighted Averaging Operators", *Systems, Man, and Cybernetics*, Part B: Cybernetics, IEEE Transactions, v.29(2), p. 141-150.
9. R.R.Yager (1987) "On the aggregation of processing units in neural networks", Proc. 1st IEEE Int. *Conference on Neural Networks*, San Diego, v.2, p. 927–933.
10. R.R.Yager (1987) "A note on weighted queries in information retrieval systems", J. Amer. Soc. Information Sciences, v.28, p. 23–24.
11. R.R.Yager (1991), "Connectives and quantifiers in fuzzy sets", *Fuzzy Sets and Systems,* v.40, p. 39-75.
12. Vicenç Torra, Lluís Godo (1997), "Averaging continuous distributions with the WOWA operator". 1997.
13. Vicenç Torra (2000), "The WOWA operator and the interpolation function W*: Chen and Otto's interpolation method revisited", *Fuzzy Sets and Systems*, v.113(3), p. 389-396.
14. Vicenç Torra (2001), "Empirical analysis to determineWeighted OWA orness", FUSION 2001: Proceedings of the Fourth International Confeence on Information Fusion, International Society of Information Fusion, V.2, p. 11-16.
15. Vicenç Torra, Yasuo Narukawa, *Modeling Decisions,* Springer-Verlag, 2007.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

ĐA TRỪU TƯỢNG KẾT HỢP TINH LỌC TRONG KIỂM TRA MÔ HÌNH HƯỚNG KÝ HIỆU

Bùi Hoài Thắng[1] và Nguyễn Thị Hồng Phấn[1]
[1] *Khoa Khoa học & Kinh tế Máy tính, Trường Đại học Bách Khoa Thành phố Hồ Chí Minh*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Multiple abstraction refinement in symbolic model checking

Từ khóa:
Tinh lọc, đa trừu tượng, kiểm tra mô hình hướng ký hiệu

Keywords:
Multiple abstraction refinement, symbolic model checking

ABSTRACT

In model checking, the state space explosion prevents verifying the large systems because exhaustive search fails to find errors. The problem becomes much more serious recently when the size and the complexity of the system-under-check are constantly growing up. Abstraction is the most effective method to overcome this problem. It is an approach used to reduce the size (of the state space) of the model in the reality. However, how to use abstraction effectively in reducing model checking efforts is still a question.

This work proposes a multiple abstraction refinement technique in symbolic model checking that allows to verify large state space systems. The proposed technique employs the abstraction refinement algorithm CEGAR proposed by Clarke et al. in 2000 combining with the multiple abstraction method of Qian and Nymeyer. Instead of checking on concrete model, an abstract model is verified first. If a counter-example is found, the searching scope will be widen on the next abstract model and so on, even on the original model. This process can be called refinement. The experimental results show that the new approach improves the performance of the model checking process.

TÓM TẮT

Trong kiểm tra mô hình, bùng nổ không gian trạng thái gây trở ngại đến việc kiểm chứng các hệ thống lớn vì quá trình tìm kiếm lỗi có thể không hoàn thành. Vấn đề càng trở nên nghiêm trọng hơn khi kích thước và sự phức tạp của hệ thống cần kiểm tra đang liên tục lớn dần lên. Trừu tượng là phương pháp hiệu quả nhất để giải quyết vấn đề này. Đây là một cách tiếp cận để giảm kích thước (của không gian trạng thái) của mô hình trong thực tế. Tuy nhiên, làm thế nào để sử dụng trừu tượng một cách hiệu quả trong việc giảm chi phí kiểm tra mô hình vẫn còn là câu hỏi.

Công trình này đề xuất kỹ thuật tinh lọc đa trừu tượng trong kiểm tra mô hình cho phép kiểm chứng các hệ thống có không gian trạng thái lớn. Phương pháp đề xuất là sự tận dụng giải thuật CEGAR được Clarke đề xuất vào năm 2000 kết hợp với phương pháp đa trừu tượng của Qian & Nymeyer. Thay vì kiểm tra trên mô hình thực, mô hình trừu tượng được kiểm chứng đầu tiên. Nếu tìm thấy counter-example, phạm vi tìm kiếm được mở rộng trên các mô hình trừu tượng tiếp theo và cứ thế, thậm chí là trên mô hình gốc. Quá trình này được gọi là tinh lọc. Kết quả thực nghiệm cho thấy rằng phương pháp mới cải thiện hiệu suất của quá trình kiểm tra mô hình.

1 GIỚI THIỆU

Trong công nghiệp, hầu hết các nhà sản xuất đều mong muốn sản phẩm của họ thâm nhập thị trường nhanh chóng với chất lượng cao. Với mục tiêu phải bảo đảm chất lượng, trong lĩnh vực công nghệ thông tin và truyền thông có nhiều phương pháp kiểm định chất lượng được áp dụng vào dây chuyền sản xuất như chứng minh đúng đắn (theorem proving), *kiểm tra mô hình* (model checking)[3] và kiểm thử (testing). Mỗi phương pháp có sức mạnh riêng và người kỹ sư hệ thống phải xem xét phương pháp nào thích hợp cho từng nhu cầu cụ thể. *Kiểm tra mô hình* được yêu thích vì sự tự động hóa và nó có thể được áp dụng vào kiểm tra cho các hệ thống phần cứng và phần mềm chưa hoàn chỉnh.

Trong kiểm tra mô hình, *mô hình* là sự đặc tả của hệ thống cần được kiểm tra, mô hình được sử dụng cho trong quá trình tìm kiếm các vi phạm các thuộc tính cần kiểm tra. Khi một vi phạm được tìm thấy, một counter-example được sinh ra để giúp các kỹ sư định vị được lỗi. Không may, khi kích thước và độ phức tạp của các hệ thống cần kiểm tra lớn liên tục, quá trình kiểm tra trở nên khó khăn hơn.

Gần đây, *kiểm tra mô hình hướng ký hiệu* trở thành hướng nghiên cứu được quan tâm và phổ biến nhất trong kiểm tra mô hình. Nó làm việc trên một tập các trạng thái và sự dịch chuyển giữa các trạng thái hơn là các yếu tố độc lập và biểu diễn các tập hợp là các biểu thức Boolean (của các biến). Vì vậy, các hệ thống có không gian trạng thái lớn có thể được kiểm chứng trên thực tế [8]. Thật ra, sự bùng nổ trạng thái vẫn xảy ra trong quá trình tìm kiếm khi kích thước của các không gian trạng thái lớn dần theo hàm mũ. Để giải quyết vấn đề bùng nổ, kỹ thuật trừu tượng có thể được dùng để giảm kích thước của mô hình.

Tuy nhiên, một phương pháp trừu tượng tốt vẫn đưa ra counter-example giả. Clarke *et al*. [4] đã đề xuất việc *tinh lọc* trên các counter-example giả có thể bị loại bỏ (abstraction refinement). Thay vì Clarke sử dụng phương pháp tinh lọc CEGAR, Qian K. *et al.* [13] giới thiệu kỹ thuật *đa trừu tượng* (multiple abstraction) để tinh lọc trừu tượng. Nếu mô hình trừu tượng không có lỗi, hệ thống được kiểm chứng là không có lỗi mà không cần kiểm tra mô hình gốc ban đầu. Dĩ nhiên, nếu tất cả các mô hình trừu tượng chứa lỗi, hệ thống thực sẽ được kiểm tra. Ngoài ra, tinh lọc trừu tượng phải giải quyết bùng nổ không gian trạng thái nếu phải thực hiện tinh lọc counter-example giả trên mô hình gốc. *Đa trừu tượng* cũng đã giải quyết cùng vấn đề này khi xem tất cả các mô hình trừu tượng và mô hình gốc độc lập mà không sử dụng lại counter-example từ các bước tìm kiếm trước đó để dẫn hướng tìm lỗi.

Trong bài báo cáo này, chúng tôi đề xuất một phương pháp tìm kiếm lỗi tận dụng ưu điểm của *tinh lọc trừu tượng*[4] và *đa trừu tượng* [3] được gọi là **MAR** (multiple abstraction refinement). Nói cách khác, giải thuật MAR là sự kết hợp của đa trừu tượng và tinh lọc trừu tượng. Hiện tại chỉ hỗ trợ kiểm tra mô hình trực tiếp (directed model checking).

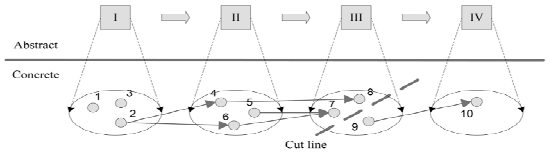
Phần còn lại của bài báo được tổ chức như sau: phần 2 trình bày về các công trình liên quan (giải thuật *CEGAR*, *đa trừu tượng* và phương pháp phân tích biến VRK). Các công trình này là tiền đề cung cấp kiến thức cho giải thuật mới được đề xuất *giải thuật đa trừu tượng kết hợp tinh lọc*, kết quả thực nghiệm cũng được trình bày trong phần này. Phần cuối trình bày kết luận và đề xuất.

2 PHƯƠNG PHÁP NGHIÊN CỨU

2.1 Công trình liên quan

2.1.1 Giải thuật CEGAR

Clarke *et al.* 2000 [4] đề xuất giải thuật tinh lọc trừu tượng dùng counter-example dẫn hướng (Counter-Example Guided Abstraction Refinement). Nếu một counter-example được tìm thấy trên mô hình trừu tượng, nó sẽ được xác nhận lại trên mô hình gốc. Đầu tiên, mô hình trừu tượng được sinh ra dựa trên mô hình gốc bằng hàm trừu tượng. Kế đó, giải thuật kiểm tra được vận dụng vào mô hình trừu tượng để tìm lỗi. Nếu không có lỗi trên mô hình trừu tượng, giải thuật dừng lại và hệ thống cần kiểm tra được trả lời là đúng với các thuộc tính mong muốn. Nếu một counter-example giả được tìm thấy, bộ kiểm chứng sẽ tinh lọc nó trên mô hình gốc. Kết quả cuối cùng sẽ được xác nhận đó có thật sự là counter-example hay không và báo cáo lên người dùng. Với giải thuật này, mức độ hiệu quả của công cụ kiểm tra phụ thuộc vào hàm trừu tượng và giải thuật tinh lọc. Phương pháp vẫn đối mặt với vấn đề như*: loop counter-example* [4], khi một counter-example trên mô hình trừu tượng có chứa vòng lặp, và *non-strong connectivity* (khi một trạng thái trừu tượng đại diện cho một tập các trạng thái thực không được kết nối với nhau). Trạng thái trừu tượng *III* trong Hình 1 là ví dụ của trường hợp thứ hai.



Hình 1: Hệ thống trừu tượng [7]

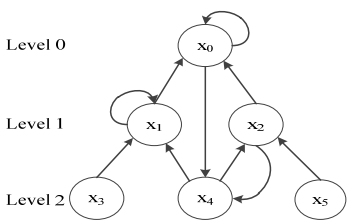
2.1.2 Giải thuật đa trừu tượng

Qian K. *et al.*[13] đề xuất chiến thuật *đa trừu tượng* để giảm chi phí kiểm tra. Dựa trên giả thuyết có các chuỗi mô hình trừu tượng $M_0 \subseteq M_1 \subseteq \ldots$ sao cho trạng thái trừu tượng S^i trên mô hình M_i được tách ra nhiều trạng thái S^{i1}, S^{i2},… trên M_{i+1} (mức độ trừu tượng của M_i nhiều hơn M_{i+1},). Giải thuật tìm kiếm chạy qua các mô hình trừu tượng theo thứ tựM_0, M_1,…tìm kiếm các vi phạm trong các mô hình trừu tượng này. Quá trình tìm kiếm có thể dừng lại nếu không tìm thấy bất kỳ counter-example nào trong bất cứ M_i. Ngược lại, quá trình tìm kiếm tiếp tục phát hiện ra lỗi trên các mô hình trừu tượng còn lại thậm chí là trên mô hình gốc. Giải thuật đa trừu tượng đối xử với tất cả các mô hình trừu tượng như các mô hình riêng rẽ và theo thứ tự. Nó chỉ xem thông tin có được về counter-example của mô hình trừu tượng trong lần tìm kiếm trước đó là heuristic. Dĩ nhiên, vấn đề các trạng thái trừu tượng có kết nối với nhau nhưng thật sự trên mô hình thật các trạng thái không có đường nối giữa chúng (non-strong connected abstract state) vẫn còn tồn tại khi kết quả có thể dẫn tới lần tìm kiếm kế tiếp bị sai đường. Trong trường hợp xấu nhất, giải thuật phải thực thi trên mô hình thật để tìm ra counter-example thật sự.

Tuy hiệu suất của giải thuật đa trừu tượng [13] rất tốt, nó vẫn cần được cải thiện để sử dụng lại thông tin từ giai đoạn tìm kiếm trước đó để chỉ tập trung vào các vùng đáng nghi ngờ trong mô hình trừu tượng kế tiếp có khả năng chứa các vi phạm.

2.1.3 Phương pháp phân tích biến VRK

Phần này giới thiệu sơ lược về phương pháp phân tích biến VRK [19]. Đây là một phương pháp để tạo ra mô hình trừu tượng. Bài báo [19] áp dụng ý tưởng giải thuật PageRank vào phân tích biến để tính độ quan trọng của biến trong đồ thị phụ thuộc biến (xem Hình 2).



Hình 2: Đồ thị phụ thuộc biến[18]

Giải thuật phân tích biến hội tụ VRK được mô tả như sau:

a. Áp dụng giải thuật PageRank để tính toán độ quan trọng của mỗi biến trong đồ thị:

$$PR(A) = \frac{(1-d)}{N} + d\left(\sum_{B \in Parent(A)} \frac{PR(B)}{N(B)}\right)$$

N là số lượng biến, *Parent(A)* là một tập tất cả các biến mà A phụ thuộc vào, *N(B)* là số lượng các biến bị phụ thuộc vào B,d là hệ số damping được thiết lập giá trị giữa 0 và 1. Theo Brin *et al.* [1], giá trị tốt nhất cho d là 0.85.

b. Điều chỉnh mức độ quan trọng của biến có được từ bước i

Dựa vào mức (level) của biến trên đồ thị, thực hiện điều chỉnh lại độ quan trong của biến theo công thức sau:

$$E(x) = E(x) + \frac{MaxLevel - Level(x)}{MaxLevel}$$

MaxLevel là giá trị lớn nhất của level có trong đồ thị, *Level(x)* là level của biến x trên đồ thị.

2.2 Giải thuật đề xuất

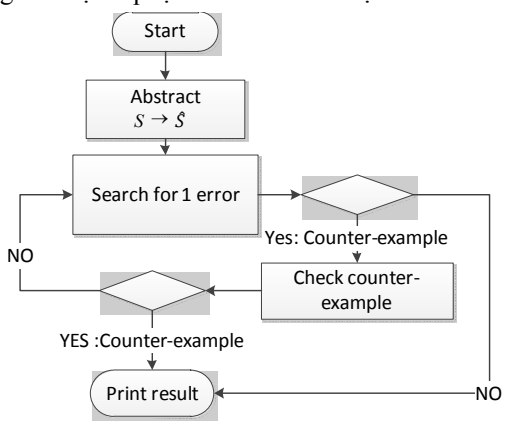
Dựa trên giải thuật *đa trừu tượng* và *CEGAR*, chúng tôi đề xuất một phương pháp mới với ưu điểm của cả hai. Với *CEGAR*, bất cứ khi nào một counter-example được tìm thấy trên các mô hình trừu tượng, nó sẽ được tinh lọc trên hệ thống gốc. Với *đa trừu tượng*, việc tìm kiếm sẽ được thực hiện trên mô hình trừu tượng kế tiếp (theo mức độ trừu tượng của các mô hình giảm dần) mỗi khi lỗi được tìm thấy trên các mô hình trừu tượng trước, quá trình tìm kiếm không sử dụng lại thông tin về lỗi ở mô hình trước, đây cũng chính là nhược điểm của giải thuật. Giải thuật mới vận dụng ý tưởng *đa trừu tượng* từ [13] và *tinh lọc* counter-example từ CEGAR [4].

Giống như *đa trừu tượng*, giải thuật mới sẽ duyệt lần lượt qua một chuỗi các mô hình trừu tượng $M_0 \subseteq M_1 \subseteq \ldots$, để tìm kiếm các vi phạm (mức độ trừu tượng M_0 là nhiều nhất). Counter-example tìm thấy trên M_0 sẽ được xác nhận lại trên M_1 và tiếp tục theo thứ tự mức độ trừu tượng M_2, M_3...và sau đó là trên hệ thống gốc. Counter-example tìm thấy trên hệ thống gốc là một chứng cứ cho việc vi phạm các thuộc tính cần kiểm tra.

Một chuỗi quá trình kiểm tra có thể được kết thúc sớm hơn nếu counter-example không được xác nhận lại trên bất cứ mô hình trừu tượng M_i nào (kể cả mô hình gốc). Trong trường hợp này, lỗi được tìm thấy trên M_0 sẽ bị từ chối (vì M_i xác nhận

không phải lỗi) và quá trình tìm kiếm trên M_0 sẽ tiếp tục.

Giải thuật được minh họa như trong Hình 3. Trong sơ đồ này, khối *Checkcounter-example* là hàm đệ qui gọi chính nó nhiều lần với mô hình trừu tượng M_i hoặc mô hình gốc lúc cuối. Nếu không có counter-example được tìm thấy trên M_0, quá trình kiểm tra kết thúc, báo cáo kết quả không có lỗi. Nếu quá trình xác nhận counter-example thất bại, giải thuật tiếp tục tìm kiếm cho một lỗi khác.



Hình 3: Quá trình tinh lọc trừu tượng

2.1 Thực nghiệm

2.1.1 Môi trường thực nghiệm

Trong thử nghiệm này, chúng tôi sử dụng lại các mô hình được cung cấp từ trang web NuSMV. Các mô hình này là các mô hình đúng được sử dụng để kiểm chứng giải thuật tìm lỗi mới. Các mô hình được sử dụng là ***brp***, ***dme***, ***gigamax***, ***msi_wtra***, ***periodic***.

Các thử nghiệm được thực hiện trên máy tính core i3-2330M 2.2 GHz với 4MB RAM chạy hệ điều hành Ubuntu 12.04 với công cụ Golfer[1]. Kết quả được được tính trung bình của các lần chạy theo đơn vị thời gian là giây.

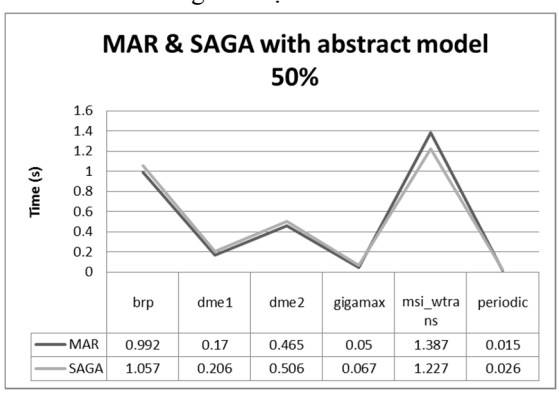
Trong thử nghiệm này, chúng tôi so sánh thời gian tìm lỗi của giải thuật mới đề xuất với giải thuật đa trừu tượng và giải thuật tìm lỗi gốc của NuSMV (giải thuật vét cạn, BFS trong bài bài toán invariant model checking). Rất tiếc là chúng tôi không thể so sánh với giải thuật CEGAR vì chúng tôi không có công cụ hiện thực của Clarke.

[1] Golfer là công trình của (Qian & Nymeyer, 2004-2005) và (Bui & Nymeyer, 2008-2009) dựa trên NuSMV.

2.1.2 Kết quả thực nghiệm

Trong phần này, chúng tôi so sánh giữa giải thuật *đa trừu tượng kết hợp tinh lọc* (gọi tắt là **MAR**) và giải thuật *đa trừu tượng* (gọi là **SAGA**) khi áp dụng cùng một phương pháp phân tích biến VRK (tham khảo[19]).

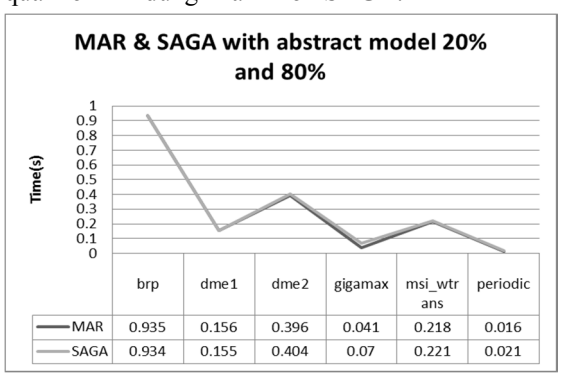
Với một mô hình trừu tượng (giữ lại 50% biến) ở Hình 4, trường hợp này giải thuật MAR chỉ dùng một mô hình trừu tượng, nên có thể xem nó tương tự như giải thuật CEGAR. Nhìn chung, giải thuật MAR nhanh hơn giải thuật SAGA.



	brp	dme1	dme2	gigamax	msi_wtrans	periodic
MAR	0.992	0.17	0.465	0.05	1.387	0.015
SAGA	1.057	0.206	0.506	0.067	1.227	0.026

Hình 4: So sánh MAR và SAGA với chỉ một mô hình trừu tượng

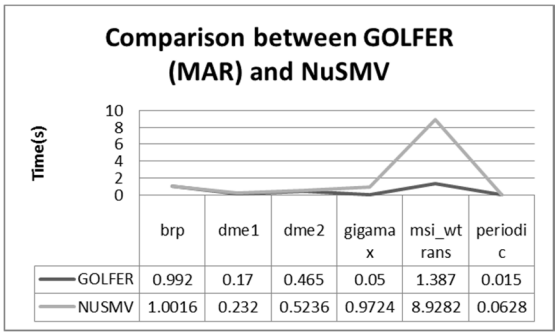
Với những ưu điểm của giải thuật *đa trừu tượng kết hợp tinh lọc*, chúng tôi có thể đặc tả một chuỗi danh sách các mô hình trừu tượng. Công cụ kiểm tra có thể ngừng lại giữa quá trình kiểm chứng mà không cần kiểm tra hết tất cả các mô hình trừu tượng và mô hình gốc. Trong thực nghiệm này, chúng tôi sử dụng hai mô hình trừu tượng được loại bỏ 80% và 20% biến dựa trên giải thuật VRK (tương ứng giữ lại 20% và 80% biến theo độ quan trọng). Kết quả cho thấy như ở Hình 5. Một lần nữa, nhìn tổng quát MAR báo cáo kết quả mô hình đúng nhanh hơn SAGA.



	brp	dme1	dme2	gigamax	msi_wtrans	periodic
MAR	0.935	0.156	0.396	0.041	0.218	0.016
SAGA	0.934	0.155	0.404	0.07	0.221	0.021

Hình 5: So sánh MAR và SAGA với hai mô hình trừu tượng

Để chắc chắn tính hữu dụng của giải thuật mới, chúng tôi cũng thực hiện so sánh nó với giải thuật NuSMV ban đầu. Kết quả ở Hình 6 chỉ ra rằng giải thuật MAR là tốt nhất.



Hình 6: So sánh MAR với giải thuật tìm kiếm NuSMV

3 KẾT LUẬN VÀ ĐỀ XUẤT

3.1 Kết luận

Công trình này giới thiệu một kỹ thuật kiểm tra mô hình dựa trên việc *đa trừu tượng kết hợp tinh lọc* counter-example được đề xuất. Kỹ thuật mới có được ưu điểm của hai kỹ thuật trước đó. Kỹ thuật mới sử dụng *đa trừu tượng* để xác nhận lại counter-example được tìm thấy trên mô hình trừu tượng nhiều nhất (mô hình trừu tượng được lược bỏ hầu hết các biến) và sử dụng ý tưởng *tinh lọc* counter-example để tinh lọc counter-example (giải thuật này cho phép tinh lọc trên nhiều mô hình trừu tượng trước khi kiểm tra trên mô hình gốc so với CEGAR chỉ sử dụng duy nhất một mô hình trừu tượng và mô hình gốc). Sự cải tiến ở đây cho phép kiểm tra trên nhiều mô hình trừu tượng có kích thước không gian trạng thái nhỏ hơn mô hình gốc và chính vì thế mà có thể giảm công sức kiểm tra.

Kết quả thử nghiệm đã khẳng định các giả thuyết. Dĩ nhiên, giải thuật của chúng tôi vẫn phải đối mặt cùng vấn đề của *CEGAR* hoặc giải thuật *đa trừu tượng* nếu mô hình gốc có chứa lỗi. Trong trường hợp đó, giải thuật sẽ vẫn phải tìm kiếm lỗi trên mô hình gốc để sinh ra counter-example.

3.2 Đề xuất

Giải thuật cần được mở rộng để hỗ trợ các hướng kiểm tra khác (thay vì hiện nay chỉ hỗ trợ kiểm tra bất biến – invariant model checking). Đây sẽ là công việc của tương lai.

LỜI CẢM TẠ

Nhóm tác giả xin chân thành cảm ơn sự hỗ trợ kinh phí của Trường Đại học Bách Khoa Thành phố Hồ Chí Minh thông qua đề tài C2013-20-07/HĐ-KHCN.

TÀI LIỆU THAM KHẢO

1. Brin, Sergey and Page, Lawrence. The anatomy of a largescale hyper-textual Web search engine. *InProc. of WWW7*, pages 107-117, Brisbane, Australia, 1998.
2. Clarke E. M., etal. Formal methods: state of the art and future directions. ACM Computing Surveys, 1996.
3. Clarke E. M., Grumberg O., and Peled D. A., *Model Checking*, MIT Press, 1999
4. Clarke E. M., Grumberg O., Jha S., Lu Y., and Veith H., Counter-example-Guided Abstraction Refinement*, CAV'00, LNCS*, vol. 1855, pp. 154-169, 2000.
5. Dennis Dams: Abstraction in Software Model Checking: Principles and Practice (Tutorial Overview and Bibliography). *SPIN 2002*: 14-21
6. Edelkamp S., Lluch-Lafuente A., and Leue S., Directed Explicit Model Checking with HSF-SPIN, *SPIN'01, LNCS*, vol. 2057, pp. 57-79, 2001.
7. Fei He, Xiaoyu Song, William N.N. Hung, Ming Gu, Jiaguang Sun, "Integrating Evolutionary Computation with Abstraction Refinement for Model Checking," *IEEE Transactions on Computers*, vol. 59, no. 1, pp. 116-126, Jan. 2010, doi:10.1109/TC.2009.105
8. J. R. Burch, E. M. Clarke, K. L. McMillan, D. L. Dill, and L. J. Hwang. 1992. Symbolic model checking: 1020 states and beyond. Inf. Comput. 98, 2 (June 1992), 142-170.
9. Menzies T., Owen, D., and Richardson J., The Strangest Thing about Software, *IEEE Computer*, vol. 40, no. 1, pp. 54-60, 2007.
10. Pelanek R., Typical structural properties of state spaces, *LNCS*, vol. 2989, pp. 5-22, 2004.
11. Qian K., Nymeyer A., Guided Invariant Model Checking Based on Abstraction and Symbolic Pattern Databases, *TACAS'04, LNCS*, vol. 2988, pp. 497-511, 2004.
12. Qian K., Nymeyer A., Abstraction-guided model checking using symbolic IDA* and heuristic synthesis, *FORTE'05, LNCS*, vol. 3731, pp. 275-289, 2005.

13. Qian K., Nymeyer A, and Susanto S.. Experiments with Multiple Abstraction Heuristics in Symbolic Verification, *SARA 2005, LNAI 3607*, pp. 290-304, 2005
14. Qing Cui, Alex Dekhtyar., On Improving Local Website Search Using Web Server Traffic Logs: *A Preliminary Report*, 2005.
15. Thang. H. Bui, Nymeyer A., Formal Verification Based on Guided Random Walks, *iFM'09, LNCS*, vol. 5423, pp. 72-87, 2009a.
16. Thang. H. Bui, Nymeyer A., Formal Model Simulation: Can it be Guided?, *SSBSE'09*, pp. 93-96, IEEE CS, 2009b.
17. Thang. H. Bui, Nymeyer A., Heuristic Sensitivity in Guided Random-Walk Based Model Checking, *SEFM'09*, pp. 125-134, IEEE CS, 2009c.
18. Thang. H. Bui, Phuong. B.K. Dang, Yet another variable dependency analysis for abstraction guided model checking, *SEATUC 2012*.
19. Thang. H. Bui, Phan. T. H. Nguyen, Convergent Propagtion variable dependency analysis for multiple abstraction model checking, *SEATUC 2013*
20. Thomas A. Henzinger, Ranjit Jhala, Rupak Majumdar and Gregoire Sutre.Lazy Abstraction. *In ACM SIGPLAN-SIGACT Conference on Principles of Programming Languages*, pages 58-70, 2002.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

NHẬN DẠNG MÃ ĐỘC SỬ DỤNG CƠ CHẾ BĂM THEO CHỈ MỤC TRÊN KHÔNG GIAN DỮ LIỆU PHÂN HOẠCH

Trương Minh Nhật Quang[1]
[1] *Khoa Công nghệ Thông tin, Trường Đại học Kỹ thuật - Công nghệ Cần Thơ*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Recognizing malicious code using hash indexing mechanism in categorical data space

Từ khóa:
Mã độc, nhận dạng mã độc, băm theo chỉ mục, không gian dữ liệu phân hoạch

Keywords:
Malicious code, recognize malicious code, hash indexing, categorical data space

ABSTRACT

To protect a computer system from threat infections, an anti-virus system needs to scan for malicious codes which may appear in target systems. In this paper, we present a technique to recognize malicious codes quickly by using hash indexing mechanism in categorical space. First, in training phase, the dataset of malicious codes will be separated into clusters which have the same characteristics. After that, we build a special rule set in a hash indexing format of ordered cluster-buckets. Next, in recognition phase, we extract object's characteristics and code them into a checksum value by using some popular hash algorithms. Then this value is used as a key to search for the same rule in indexed rule space. Finally, the system returns the scanning process results.

We have built this technique for D2 Anti-virus 2013 running on Windows XP SP3 in a computer with Intel Core 2 Duo E7200 – 2.53 GHz. Using a dataset of 615,880 malicious signatures, D2 needs only 05 seconds to examine 105,330 MB of 8,696 executable files. Its average scanning speed is about 21,651MBps. The experimental results denote that this is an effective technique to improve the scanning speed of anti-virus systems nowadays.*

TÓM TẮT

Để bảo vệ máy tính khỏi các đe dọa lây nhiễm, hệ phòng chống virus máy tính cần quét kiểm tra mã độc trong hệ thống đích. Trong bài viết này, chúng tôi trình bày kỹ thuật nhận dạng nhanh mã độc sử dụng cơ chế băm theo chỉ mục trên không gian phân hoạch. Đầu tiên, trong giai đoạn luyện, tập mẫu chữ ký mã độc được phân thành các cụm có cùng đặc điểm. Sau đó, chúng tôi xây dựng một tập luật đặc biệt dưới dạng bảng băm các bucket luật được sắp xếp thứ tự theo cụm. Tiếp theo, ở giai đoạn nhận dạng, chúng tôi tiến hành trích chọn đặc trưng và biến đổi thành một giá trị tổng kiểm đại diện cho đối tượng bằng các thuật giải băm phổ biến. Giá trị này sau đó được dùng làm khóa tìm kiếm luật nhận dạng của đối tượng trong không gian luật đã được sắp xếp. Cuối cùng, hệ trả về kết quả quá trình duyệt quét.

Chúng tôi đã cài đặt kỹ thuật này cho hệ D2 Anti-virus 2013 chạy hệ điều hành Windows XP SP3 trên máy tính Intel Core 2 Duo E7200 – 2.53 GHz. Sử dụng tập 615,880 mẫu mã độc, D2 chỉ tốn 5 giây để kiểm tra 105,330 MB dữ liệu của 8,696 tập tin thực thi. Tốc độ quét trung bình của D2 đạt 21,651 MB/giây. Kết quả thực nghiệm chứng tỏ đây là kỹ thuật hiệu quả nhằm tăng tốc duyệt quét cho các hệ phòng chống virus máy tính ngày nay.*

1 GIỚI THIỆU

Trong bối cảnh ngày càng gia tăng các cuộc tấn công thâm nhập mạng, các hệ phòng chống virus máy tính không ngừng cải tiến nhằm sớm phát hiện các loại mã độc hại, kịp thời loại trừ các tác nhân thâm nhập, tăng cường công tác bảo vệ an toàn hệ thống và an ninh mạng. Hiện nay, hiệu quả hoạt động của các hệ phòng chống virus máy tính rất được người dùng quan tâm. Để cải tiến tốc độ truy vấn chữ ký mã độc trên cơ sở dữ liệu (CSDL) lớn, chúng tôi nghiên cứu kỹ thuật nhận dạng nhanh mã độc sử dụng cơ chế băm theo chỉ mục trên không gian phân hoạch. Đầu tiên, tập mẫu chữ ký mã độc được phân thành các cụm có cùng đặc điểm. Sau đó, chúng tôi xây dựng tập luật dưới dạng bảng băm các bucket được sắp xếp thứ tự theo cụm. Tiếp theo, chúng tôi tiến hành trích chọn đặc trưng và biến đổi thành một giá trị số đại diện cho đối tượng bằng các hàm tổng kiểm (checksum) phổ biến. Giá trị này được dùng làm khóa tìm kiếm luật nhận dạng của đối tượng trong không gian luật đã sắp xếp. Kết quả truy vấn sẽ cho biết tình trạng an ninh của đối tượng chẩn đoán có phải là mã độc hay không.

Để đánh giá hiệu quả kỹ thuật, chúng tôi xây dựng tập 615,880 mẫu mã độc và động cơ chẩn đoán của hệ D2 Anti-virus* 2013 (phiên bản dành cho Windows XP). Máy tính sử dụng bộ vi xử lý Intel Core 2 Duo E7200 – 2.53GHz. Kiểm tra 8,696 tập tin thực thi, hệ D2 tốn 5 giây, tốc độ quét trung bình 21,651MB/giây. Kết quả thực nghiệm chứng tỏ đây là kỹ thuật hiệu quả tăng tốc duyệt quét cho các hệ phòng chống virus máy tính ngày nay.

2 CÁC VẤN ĐỀ LIÊN QUAN

2.1 Mã độc trong tấn công an ninh mạng

2.1.1 Vai trò của mã độc trong kịch bản tấn công an ninh mạng

Trong các cuộc tấn công mạng, đặc biệt là hình thức tấn công từ chối dịch vụ lan tràn (DDoS – Distributed Denial of Service), hacker thường cài đặt các đoạn mã độc hại (malicious code) vào các phần mềm ác ý (malware) rồi tìm cách cấy vào các máy trạm (zombie) để hình thành mạng botnet, chuẩn bị cho cuộc tấn công. Dưới sự điều khiển của hacker, zombie sẽ thu thập, đánh cắp thông tin, liên lạc hacker... Khi lượng zombie đủ lớn, hacker sẽ phát lệnh điều khiển botnet đồng loạt tấn công vào một mục tiêu trên mạng [3].

Có nhiều dạng thức thi hành mã lệnh. Dạng mã lệnh thực thi thuận tiện cho việc truyền tải, thi hành trên máy đích hiện nay là định dạng tập thực thi (executable files). Mỗi mã độc có đoạn mã đặc trưng nhận dạng gọi là chữ ký (signature) mã độc [5]. Có hai hình thức thiết kế mã độc: thi hành phụ thuộc và thi hành độc lập.

2.1.2 Mã độc thi hành phụ thuộc

Tiêu biểu cho dạng mã độc này là các loại virus máy tính (gọi tắt là virus), trojan horse (gọi tắt là trojan), spyware, adware... Virus là loại mã độc có khả năng ký sinh mã vào các tập thực thi khác và nắm quyền thực thi khi ứng dụng chủ thi hành. Mã virus được thiết kế bằng Hợp ngữ (Assembly Language), độc lập với ngôn ngữ lập trình của vật chủ. Các thủ tục thi hành của trojan (và spyware, adware...) được hacker cài đặt trong lúc thiết kế mã độc. Virus ký sinh vào các tập thực thi tạo hiệu ứng lây lan, trong khi trojan (và spyware, adware...) không tự lây lan; chúng được ngụy trang dưới vỏ bọc các phần mềm hiền lành, lừa người sử dụng tải về máy và thi hành chúng [8].

2.1.3 Mã độc thi hành độc lập

Việc cài đặt malware vào máy người dùng dưới dạng một phần mềm hoàn chỉnh có thể gây nhiều bất lợi: kích thước lớn, thay đổi registry, lộ diện tiến trình... Để khắc phục, hacker biên dịch mã độc dưới dạng file thực thi rút gọn rồi lan truyền trên mạng thông qua các lỗ hổng bảo mật hệ thống. Tiêu biểu cho dạng mã độc này là các loại sâu mạng (networm), cửa hậu (backdoor), intruder, dropper, rootkit... Chúng được hacker ưa chuộng vì kích thước nhỏ, lây lan nhanh, hoạt động độc lập (stand alone), dễ ẩn náu [4].

Hiệu quả của kế hoạch tấn công mạng phụ thuộc vào số lượng zombie. Để gia tăng khả năng lây lan, kéo dài thời gian ẩn náu..., hacker thường kết hợp nhiều hình thức mã độc: virus chứa sâu mạng, trojan kết xuất dropper... Do đó, phân loại mã độc như trên chỉ là tương đối.

2.2 Các hệ phòng chống virus máy tính

Lịch sử phát triển mã độc bắt nguồn từ virus máy tính (sau đây gọi tắt là virus). Thời kỳ đầu các phần mềm phòng chống virus được gọi là anti-virus. Ngày nay thuật ngữ anti-virus (AV) dùng để chỉ loại phần mềm bảo vệ máy tính khỏi sự xâm nhập của các loại mã độc hại (virus, trojan, worm, backdoor, rootkit, spyware...) bằng cách đối chiếu dữ liệu hệ thống với thông tin mô tả các loại mã độc đã biết trong một CSDL được cập nhật thường xuyên [8]. Các AV hoạt động dựa vào tập mẫu thường áp dụng các tiếp cận chuỗi mã, tiếp cận hành vi, tiếp cận máy học.

2.2.1 Tiếp cận chuỗi mã

Hoạt động theo nguyên lý so khớp mẫu (signature matching), các AV đối chiếu thông tin thi hành, mã lệnh hoặc/và dữ liệu của đối tượng chẩn đoán với thông tin mã độc được tổ chức, cập nhật trong CSDL [7]. Tiếp cận chuỗi mã giúp AV nhận dạng mã độc đã biết với độ chính xác cao. Tuy nhiên hoạt động của AV sẽ kém hiệu quả khi kiểm tra các mã độc chưa được cập nhật thông tin vào CSDL mẫu.

2.2.2 Tiếp cận hành vi

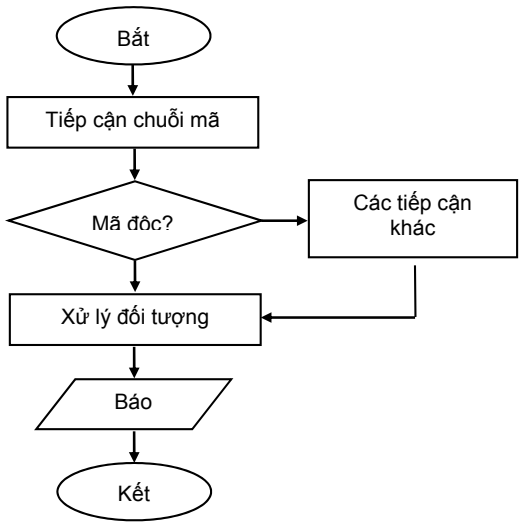
Các mã độc giống nhau thường có các hành vi giống nhau. Nghiên cứu mã độc dưới góc độ thi hành, tiếp cận này dựa vào khái niệm *hành vi* (behavior checking) để nhận dạng mã độc bằng cách tải và thử thi hành (heuristic) mã lệnh nghi ngờ trong môi trường mô phỏng (emulated environment) nhằm giải mã các hành vi lạ [7]. Tiếp cận hành vi (còn gọi là tiếp cận heuristic) giải quyết khá tốt các loại mã độc đa hình (polymorphism). Tuy nhiên khi xử lý các hành vi nghi ngờ của đối tượng, các AV heuristic cần hỗ trợ quyết định nên thường gây lúng túng cho người dùng ít kinh nghiệm.

2.2.3 Tiếp cận máy học

Tiếp cận máy học nhận dạng virus máy tính (Machine Learning Approach to Anti-virus System) xây dựng các mô hình học dựa vào cơ sở tri thức (CSTT) gồm tập mẫu chữ ký mã độc và luật nhận dạng. Tiếp cận này chia bài toán chẩn đoán mã độc thành nhiều lớp bài toán con rồi áp dụng các hình thức học phù hợp cho từng bài toán cụ thể [2][10][11]. Sử dụng kiến thức chuyên gia tổ chức trong CSTT, các hệ học 'dạy' cho máy tính xử lý các trường hợp tương tự bằng cách khái quát hóa các tình huống chẩn đoán thông qua các luật dẫn xuất (production rule) suy luận từ CSTT của hệ. Tiếp cận máy học kết hợp hệ chuyên gia giúp hệ thường xuyên cập nhật và tự tăng trưởng tri thức. Nhận dạng hướng luật giúp AV có khả năng dự báo các mẫu mã độc tương đồng. Tuy nhiên, hiệu quả chẩn đoán của AV phụ thuộc vào chất lượng CSTT của hệ.

2.2.4 Vai trò của tiếp cận chuỗi mã trong kiến trúc anti-virus

Dù áp dụng tiếp cận nào, nhiệm vụ của các AV là sớm đưa ra bằng chứng xuất hiện nếu có của một loại mã độc đã biết với tên gọi cụ thể. Vấn đề này chỉ có thể giải quyết bằng tiếp cận chuỗi mã ở giai đoạn tiền xử lý (Hình 1).



Hình 1: Lưu đồ xử lý tổng quát của AV

2.3 Cơ chế tìm kiếm, so khớp mẫu

Quét virus là quá trình tìm kiếm chữ ký mã độc trong đối tượng. Liên quan đến chủ đề bài viết có các kỹ thuật tìm kiếm tuyến tính, tìm kiếm bằng chỉ mục và tìm kiếm bằng bảng băm.

2.3.1 Tìm kiếm tuyến tính

Tìm kiếm tuyến tính (linear search) là kỹ thuật duyệt từng phần tử trong danh sách đến khi tìm thấy (so khớp) giá trị mong muốn hoặc đến cuối danh sách. Đơn giản trong thiết kế, tìm kiếm tuyến tính thích hợp cho các danh sách đủ nhỏ, danh sách chưa sắp thứ tự, các tổ chức lưu trữ tuần tự như danh sách liên kết, tập tin định kiểu… Tuy nhiên khi hoạt động trên danh sách đã sắp xếp hoặc danh sách lớn, phương pháp này không tối ưu và kém hiệu quả.

2.3.2 Tìm kiếm theo chỉ mục

Trong các hệ quản trị CSDL quan hệ, tìm kiếm theo chỉ mục (index search) là kỹ thuật tìm kiếm vị trí mẫu tin trên cấu trúc bảng dựa vào thông tin chỉ mục. Đầu tiên, thông tin đối tượng được tổ chức trong bảng 2 chiều; mỗi bảng có 1 trường khóa chính (primary key) lưu trị thuộc tính duy nhất (unique) của đối tượng. Tiếp theo, một bảng chỉ mục dạng *<primary key, #record>* được khởi tạo (*#record* chứa số hiệu mẫu tin của đối tượng trong bảng chính). Sau đó, bảng chỉ mục được sắp xếp theo (sort by) trường *primary key* rồi lưu vào tổ chức đĩa. Quá trình tìm kiếm mẫu tin có khóa *primary key* trong bảng chính được thực hiện bằng thuật toán tìm kiếm trên danh sách thứ tự (chẳng

hạn, tìm kiếm nhị phân), lấy giá trị *#record* rồi mở bảng chính, dịch đến vị trí mẫu tin tương ứng.

Sử dụng thuật toán tìm kiếm trên danh sách thứ tự nên tìm kiếm theo chỉ mục nhanh hơn tìm kiếm tuyến tính. Tuy nhiên, số mẫu tin trong bảng chính luôn bằng với số mẫu tin trong bảng chỉ mục nên kỹ thuật tìm kiếm theo chỉ mục nói chung không giảm không gian dữ liệu. Mục 2.3.3 sau đây sẽ trình bày kỹ thuật tìm kiếm hướng giảm không gian dữ liệu.

2.3.3 Tìm kiếm bằng bảng băm

Tìm kiếm bằng bảng băm (hash table search) là kỹ thuật tìm kiếm theo hướng giảm không gian dữ liệu. Ở giai đoạn thiết kế, một hàm băm (hash function) được sử dụng để đặc tả vị trí mẫu tin, sau đó lưu từng mẫu tin vào cấu trúc bảng theo mục (bucket) do hàm băm cung cấp. Ở giai đoạn truy xuất, hàm băm sẽ giúp định vị bucket phục vụ các thao tác cơ bản trên CSDL.

Hiệu quả tìm kiếm dữ liệu trên bảng băm phụ thuộc vào chất lượng hàm băm. Một hàm băm lý tưởng sẽ tạo k ánh xạ 1-1 từ tập mẫu sang bảng băm. Nếu băm 'không mịn', bảng băm thưa gây lãng phí ô nhớ (có nhiều bucket rỗi không dùng đến). Ngược lại, băm 'quá mịn' sẽ gây xung đột địa chỉ mẫu tin (hàm băm trả về một bucket cho nhiều mẫu tin khác nhau). Để giải quyết đụng độ trên bảng băm, chúng tôi sử dụng phương pháp kết nối dạng *<bucket, record list>*; trong đó, *record list* là danh sách liên kết các mẫu tin đụng độ *bucket* do hàm băm trả về.

2.4 Các nghiên cứu liên quan

2.4.1 Tình hình nghiên cứu cải tiến tốc độ quét AV trên máy cá nhân

Có nhiều yếu tố ảnh hưởng tốc độ quét của AV: tài nguyên khả dụng của hệ thống, mức ưu tiên tiến trình, kỹ thuật xử lý, tối ưu mã lệnh, tình trạng dữ liệu, số lượng mẫu trong CSDL... Đề tài này tập trung nghiên cứu các vấn đề liên quan đến dữ liệu ở hai khía cạnh: dữ liệu bên ngoài (dữ liệu chẩn đoán) và dữ liệu bên trong (CSDL mẫu của hệ).

Phân loại dữ liệu chẩn đoán, giảm không gian tìm kiếm để tránh quá tải là giải pháp được các AV sử dụng phổ biến [13]. Bên cạnh đó, một số giải pháp tăng tốc duyệt quét như lập hồ sơ tiền sử (history) lưu vết trạng thái hoạt động; ghi nhận dữ liệu an toàn [12]; đánh dấu dữ liệu nguyên trạng (not modified) [14]; tổ chức CSDL mẫu hợp lý [15]; kiểm soát các hoạt động thay đổi tập thi hành [17]; phân luồng xử lý (thread processing) [16]... cũng được các AV đầu tư nghiên cứu.

2.4.2 Tình hình nghiên cứu cải tiến tốc độ quét mã độc trên mạng theo hướng băm

Số lượng mã độc trên thị trường ngày càng nhiều khiến tập mẫu của AV gia tăng từ vài trăm ngàn đến hàng triệu chữ ký. Giải pháp cải tiến tốc độ quét mã độc theo hướng băm gần đây cũng được quan tâm nghiên cứu. Năm 2007, Ozgun Erdogan và Pei Cao đề xuất kỹ thuật Hash-AV sử dụng cơ chế đệm L2 trong CPU để quét nhanh chữ ký mã độc lan truyền trên mạng [7]. Thực nghiệm trên máy tính CPU Pentium 4 2.4GHz, sử dụng tập mẫu 120,000 chữ ký mã độc, kết quả Hash-AV cho phép tốc độ quét thông lượng trên mạng đạt 200Mb/s.

Năm 2011, Po-Ching Lin và *ctv* giới thiệu kỹ thuật Hybrid Algorithm of Backward Hashing And Automaton Tracking. Băm ngược chuỗi mẫu, ghi vết quá trình tìm kiếm, bỏ qua các chuỗi đã đối chiếu nhằm kéo dài vectơ khoảng cách trong thuật toán tìm kiếm của Aho-Corasick, kỹ thuật này giúp cải tiến 50% tốc độ thông lượng quét mạng bằng Hash-AV sử dụng CSDL mẫu của phần mềm nguồn mở Clam-AV [9].

3 NHẬN DẠNG MÃ ĐỘC SỬ DỤNG CƠ CHẾ BĂM THEO CHỈ MỤC TRÊN KHÔNG GIAN PHÂN HOẠCH

3.1 Tổ chức cơ sở tri thức

Trong các hệ học, CSTT chứa thông tin mô tả đối tượng và các luật nhận dạng đối tượng. Tiếp cận máy học và hệ chuyên gia tổ chức CSTT ở giai đoạn luyện qua bốn bước: (1) xây dựng tập mẫu, (2) phân hoạch tập mẫu, (3) rút luật nhận dạng và (4), tổ chức tập luật [2].

3.1.1 Xây dựng tập mẫu

Có nhiều dạng thức thi hành mã lệnh (và cũng từng ấy các mô tả mã độc tương ứng). Bài viết này tập trung vào các loại mã độc dạng thực thi EXE. Đầu tiên, các tập mã độc được thu thập cho chuyên gia phân tích kỹ thuật, phân loại. Sau đó, chuyên gia sẽ cập nhật các mẫu tin mô tả mã độc vào cấu trúc bảng dạng:

Malware = <Tên gọi, Tập thuộc tính, Chữ ký>

Trong đó:

- Tên gọi: tên mã độc, dùng cho giai đoạn báo cáo.
- Tập thuộc tính: các đặc trưng thi hành của mã độc, dùng phân hoạch tập mẫu.

– Chữ ký: đặc trưng 'nhân dạng' mã độc, dùng xây dựng luật nhận dạng.

3.1.2 Phân hoạch tập mẫu

Các phương pháp tìm kiếm tuyến tính trên tập mẫu lớn có nhiều hạn chế. Chúng tôi giải quyết vấn đề này bằng chiến lược tìm kiếm trên không gian tập mẫu được phân cụm.

Phân cụm (clustering) là một trong những phương pháp khai phá dữ liệu phổ biến trong các hệ học. Bằng các cấu trúc dữ liệu đặc biệt, quá trình phân cụm sử dụng các giải thuật thích hợp nhằm phát hiện sự giống nhau giữa các mục dữ liệu để gom (hoặc tách) thành từng cụm có cùng tính chất. Để phân hoạch không gian tập mẫu, chúng tôi chọn hình thức phân cụm phân cấp (*hierarchical clustering*) dạng tách nhóm. Không cần xác định số cụm từ đầu, phương pháp này sử dụng đồ thị cây các cụm (*dendogram*) với các nút biểu diễn cụm trung gian, nút lá biểu diễn các cụm kết quả.

Cây là mô hình thích hợp cho việc tổ chức dữ liệu phân cụm tách nhóm [1]. Chúng tôi chọn V-Tree để phân cụm tập mẫu [6]. Cho số nguyên m>1, cây m-phân V-Tree là một cây có:

– Các nút lá đều có số mức như nhau.

– Mỗi nút trung gian N có m/2 đến m nút con, nút gốc có từ 2 đến m con.

– Mỗi nút có nhãn chứa giá trị của nút và danh sách các nút thuộc cây con của nút.

Chúng tôi cài đặt mỗi nút V-Tree có 3 trường: (1) nhãn nút chứa trị thuộc tính, (2) danh sách các nút con và (3), danh sách các mẫu trong cụm (dùng lưu trữ thông tin chỉ mục dành cho giai đoạn xử lý, xem mục 3.1.4).

CSDL mẫu được tổ chức dưới dạng bảng 2 chiều *X(p,k)* có p cột (thuộc tính) k dòng (mẫu tin). Giải thuật phân cụm tập mẫu như sau:

```
Khởi tạo nút gốc
Đối với mỗi mẫu tin thứ j của X:
    Chọn nút gốc là Nút_hiện_tại
    Đối với mỗi thuộc tính thứ i:
        Nếu Nút_hiện_ tại chưa có nút con nhãn
        x[i,j] thì:
            Tạo nút mới có nhãn x[i,j]
            Cho nút nhãn x[i,j] là con của Nút_hiện
            tại
        Đặt nút con nhãn x[i,j] của Nút_hiện_tai làm
        Nút_hiện_tại
        Bổ sung mẫu tin j vào danh sách các mẫu
        thuộc cụm của Nút_hiện_tại
```

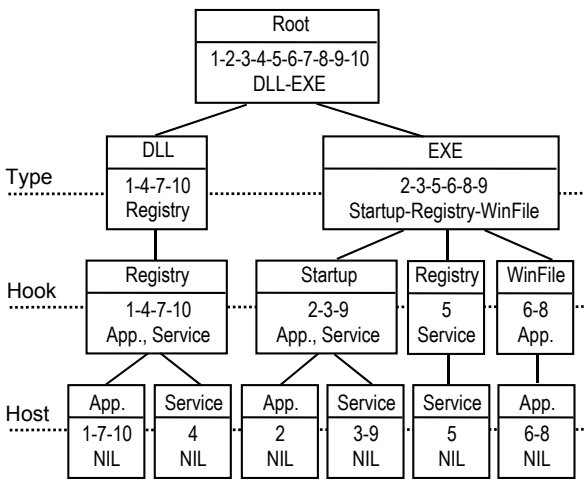
Xét CSDL mẫu gồm 5 thuộc tính (Bảng 1). Trong đó, các thuộc tính Type, Hook, Host được chọn tham gia phân cụm (p'= p–2).

Bảng 1: CSDL mẫu chứa 10 mã độc

Name	Type	Hook	Host	Signature
M1	DLL	Registry	App.	3F2D9A..
M2	EXE	StartUp	App.	E9124F..
M3	EXE	StartUp	Service	60E89F..
M4	DLL	Registry	Service	583E6F..
M5	EXE	Registry	Service	E8558B..
M6	EXE	WinFile	App.	C0317F..
M7	DLL	Registry	App.	902D3F..
M8	EXE	WinFile	App.	7E5D9B..
M9	EXE	StartUp	Service	347A9C..
M10	DLL	Registry	App.	4B80FF..

Cluster1: M1, M7, M10
Cluster2: M4
Cluster3: M5
Cluster4: M2
Cluster5: M3, M9
Cluster6: M6, M8

Hình 2: mô tả kết quả phân cụm CSDL dựng bằng V-tree, kết quả thu được 6 nhóm mã độc



Hình 3: Kết quả phân cụm tập mẫu bằng V-tree

3.1.3 Rút luật nhận dạng

Sau khi phân hoạch tập mẫu, mỗi nhánh trên V-Tree sẽ chứa một luật theo mệnh đề Horn. Rút luật nhận dạng bằng cách duyệt đường đi từ nút gốc đến nút lá. Trong ví dụ trên, tập 10 mẫu tin được tách thành 6 cụm mã độc tương ứng với 6 luật nhận dạng như sau:

1. IF (Type=DLL) AND
   (Hook=Registry) AND
   (Host=App.) THEN Cluster1
2. IF (Type=DLL) AND

(Hook=Registry) AND
(Host=Service) THEN Cluster2
3. IF (Type=EXE) AND
(Hook=Startup) AND
(Host=App.) THEN Cluster3
4. IF (Type=EXE) AND
(Hook=Startup) AND
(Host=Sevice) THEN Cluster4
5. IF (Type=EXE) AND
(Hook=Registry) AND
(Host=Service) THEN Cluster5
6. IF (Type=EXE) AND
(Hook=WinFile) AND
(Host=App.) THEN Cluster6

3.1.4 Tổ chức tập luật băm theo chỉ mục

Mục đích của phân cụm V-Tree nhằm tách tập mẫu thành các nhóm có cùng đặc tính dữ liệu về mặt *giá trị*. Các cụm kết quả (bố trí ở nút lá của V-Tree) được tách rời và không phủ nhau nên số cụm thu được (m cụm) luôn nhỏ hơn số mẫu dữ liệu (k mẫu). Điều này luôn đúng vì luật có tính khái quát, do đó chỉ cần lập luận trên m luật là có thể bao quát toàn bộ k mẫu (m<k). Như vậy, ngoài chức năng cung cấp tri thức mã độc của chuyên gia, luật nhận dạng còn có thể sử dụng làm chỉ dẫn truy xuất dữ liệu. Chúng tôi bổ sung trường thứ ba vào mỗi nút trên V-Tree nhằm cung cấp giá trị trả về cho hàm băm $H(f(A),k)$ trong phép trích chọn đặc trưng thi hành $f(A)$ của đối tượng.

Đầu tiên, tập luật nhận dạng được tổ chức dạng bảng 2 chiều $R(\psi, £, r)$ có r phần tử luật; mỗi luật ψ liên kết với danh sách £ các thể hiện thỏa luật ψ. Trong ví dụ trên, CSDL luật R có 6 phần tử luật ($r = 6$) được minh họa như sau:

Luật 1: ψ_1 » {1,7,10}
Luật 2: ψ_2 » {4}
Luật 3: ψ_3 » {5}
Luật 4: ψ_4 » {2}
Luật 5: ψ_5 » {3,9}
Luật 6: ψ_6 » {6,8}

Ký hiệu '»' đặc tả phép 'liên kết' (link-to).

Tiếp theo, các bucket luật thành viên ψ_i được biến đổi thành các giá trị tổng kiểm bằng các thuật toán băm thông dụng như CRC (Cyclic Redundancy Check), MD5 (Message - Digest Algorithm 5) hoặc SHA-1 (Secure Hash Algorithm). Sau cùng, sắp xếp tập luật theo chỉ mục checksum (Bảng 2) rồi lưu trữ CSDL luật. Kết thúc giai đoạn luyện trên máy chuyên gia.

Bảng 2: CSDL luật chứa 6 luật thành viên

No.	Rule	Checksum	Link-to
1	ψ_2	58,487,876	4
2	ψ_5	76,455,645	3-9
3	ψ_4	156,496,857	2
4	ψ_1	325,474,326	1-7-10
5	ψ_6	437,665,473	6-8
6	ψ_3	758,355,475	5

3.2 Nhận dạng mã độc dựa vào tri thức

Giai đoạn nhận dạng - xử lý trên máy khách vận dụng các thủ tục suy diễn và lập luận trong động cơ quét (scan engine), căn cứ vào thông tin trong tập chữ ký mẫu và các tri thức mô tả trong CSDL luật để chẩn đoán đối tượng A. Quá trình này được thực hiện qua bốn bước: (1) trích chọn đặc trưng, (2) mã hóa nhân dạng, (3) truy vấn luật và (4), so khớp mẫu.

3.2.1 Trích chọn đặc trưng

Đầu vào của bước này là đối tượng A cần chẩn đoán. Đầu ra là một tập mô tả các đặc trưng thi hành của đối tượng dùng cho bước mã hóa nhân dạng tiếp theo. Để tương thích với khuôn dạng tri thức mô tả mã độc M ở giai đoạn luyện, bước này sử dụng lại hàm trích chọn đặc trưng $f(M)$ của chuyên gia dạng:

$$f(A) = \{\delta_i\} \ \forall \ (i < p)$$

Trong đó, $\delta(q)$ là phép trích chọn tự động tập thuộc tính mã độc ở giai đoạn trước. Ví dụ, áp dụng hàm f trích chọn đặc trưng trên các đối tượng bất kỳ U, V nhận được kết quả như sau:

$f(U)$ = {DLL, Registry, Driver}
$f(V)$ = {EXE, Startup, App.}

Ví dụ này sử dụng $i = 3$ ($i = p' < p$)

3.2.2 Mã hóa nhân dạng

Nhiệm vụ của bước trích chọn đặc trưng là tạo lập nhân dạng (identification) của đối tượng theo khuôn dạng luật trong CSTT. Khái niệm nhân dạng trong tiếp cận học chẩn đoán mã độc dựa vào lập luận: "nếu đối tượng có nhân dạng giống với nhân dạng tội phạm thì có thể đối tượng chính là tội phạm đang truy nã".

Sau khi trích chọn đặc trưng, nhân dạng đối tượng sẽ được mã hóa bằng thuật toán băm h của chuyên gia. Ví dụ, hồ sơ chẩn đoán các đối tượng

U, *V* nhận được mã số nhân dạng sau:

$ID(U) = h(f(U)) = 87,954,543$

$ID(V) = h(f(V)) = 758,355,475$

Các mã số này sau đó được sử dụng làm giá trị tìm kiếm theo trường khóa chính Checksum trong CSDL luật của hệ.

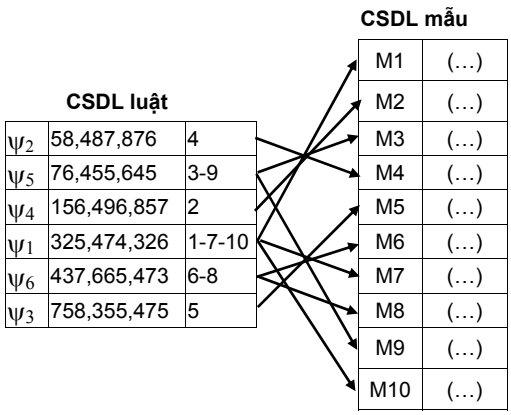
3.2.3 Truy vấn luật

Khi tra cứu mã số nhân dạng *ID*(*A*) trong CSDL luật, nếu không tìm thấy có thể kết luận *A* là đối tượng an toàn vì trong CSTT chưa có các mô tả mã độc nào có đặc trưng thi hành giống với *A*. Nếu tìm thấy, cần đưa *A* vào diện nghi vấn, sau đó tiếp tục xác minh các đặc điểm nhân dạng đặc thù để có kết luận chính xác. Trong ví dụ trên, *U* là an toàn vì trong tập luật không có giá trị checksum 87,954,543 nào. Tuy nhiên cần cảnh giác *V* vì mã số nhân dạng 758,355,475 của nó đang có trong CSDL luật.

Mỗi luật nhận dạng đại diện cho một cụm (nhóm) các mã độc cùng đặc trưng. Xác minh hồ sơ là quá trình đối chiếu thông tin nhân dạng của *A* với các mã độc cùng nhóm. Cho các số nguyên *i*, *j*; hàm băm *H* truy vấn danh sách mã độc cùng nhóm trên tập mẫu *X(p,k)* có dạng:

$$H(\psi_m) = \psi_i \gg \{X(p, j)\} \ \forall \ (i < m); (j < k)$$

Theo thiết kế này, hàm *H* chỉ đơn giản trả về danh sách ở trường thứ ba của nút lá trên nhánh V-Tree của luật ψ_i tương ứng. Chiến lược sử dụng tập luật nhận dạng làm chỉ mục băm tìm kiếm chữ ký mã độc được minh họa ở Hình 3.



Hình 3: Cơ chế băm tập luật theo chỉ mục

3.2.4 So khớp mẫu

Đây là công đoạn rà soát, đối chiếu thông tin lần cuối trước khi quyết định 'bắt giam' đối tượng nhằm hạn chế sai sót trong tác nghiệp. Có nhiều cách so khớp mẫu: dò tìm mã phổ biến, trích xuất các giá trị đặc trưng tại các địa chỉ nhạy cảm, trích chọn mã ngẫu nhiên…[5].

Trong bài viết này, so khớp mẫu là bước kiểm tra sự xuất hiện giá trị thuộc tính thứ *p* của quan hệ *X*(*p*,*k*) trong đối tượng chẩn đoán. Nếu đối chiếu đúng, có thể kết luận đối tượng *A* là mã độc. Ngược lại, tiếp tục đối chiếu các mẫu còn lại cho đến khi phát hiện. Gọi *M*(*p*,*n*) là tập mã độc có mã nhận dạng trùng với mã nhân dạng của *A*. Giải thuật so khớp mẫu như sau:

```
Tìm thấy ← false
j ← 1
Lặp lại
    Nếu Chữ ký(A) = M(p,j) thì
       Kết luận ('Phát hiện mã độc', M(1,j))
       Tìm thấy ← true;
    Ngược lại
       j ← j+1
Đến khi Tìm thấy hoặc (j>n)
```

4 KẾT QUẢ THỰC NGHIỆM

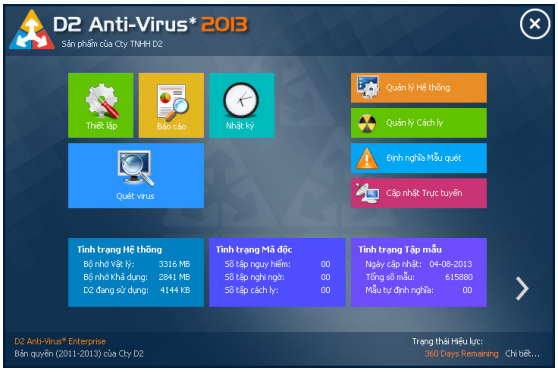
4.1 Thử nghiệm với D2 Anti-virus*

*4.1.1 Giới thiệu hệ D2 Anti-virus**

D2 Anti-virus* (Diagnose and Destroy Computer Viruses) là phần mềm quét virus hướng tiếp cận máy học của Việt Nam. Thiết kế theo mô hình của một hệ chuyên gia, hoạt động của hệ D2 gồm ba giai đoạn: giao tiếp chuyên gia (giai đoạn học), động cơ suy diễn (giai đoạn nhận dạng, xử lý) và giao tiếp người dùng (giai đoạn tổng kết, báo cáo).

Thực hiện trên máy chuyên gia, giai đoạn học xây dựng CSTT dưới dạng tập các mô tả tri thức mã độc và luật nhận dạng khẳng định dương trên tập mẫu. Giai đoạn xử lý và báo cáo thực hiện trên máy khách. Hoạt động hướng tác tử [10], động cơ quét của hệ chứa các thuật toán học phân loại, động cơ suy diễn, thực thi các lớp bài toán học và xử lý các trường hợp lây nhiễm. Cuối cùng, giai đoạn báo cáo tổng hợp kết quả các bài toán, phân tích quá trình, tạo các giao tiếp hội thoại, thu nhận ý kiến người dùng, báo cáo kết quả và tăng trưởng tri thức.

Phiên bản D2 Anti-virus* 2013 (Hình 4) có một số cải tiến quan trọng về giao diện, tăng tốc độ quét, dự báo mã độc hướng heuristic, ước lượng mã tương đồng, phát hiện hành vi lây nhiễm trên thiết bị lưu trữ cá nhân…



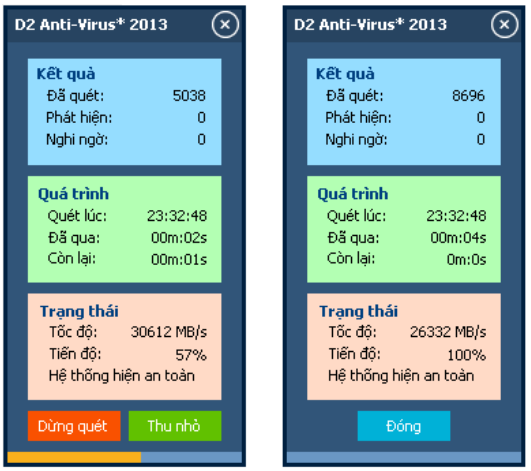
Hình 4: Phần mềm D2 Anti-Virus 2013*

*4.1.2 Kết quả thử nghiệm D2 Anti-virus**

Trong giai đoạn học, chúng tôi thiết kế gói công cụ D2 Expert Utilities hỗ trợ phân tích kỹ thuật tập mẫu, đặt tên mã độc, trích chọn đặc trưng thi hành, xây dựng CSDL, phân hoạch tập mẫu, rút luật nhận dạng, mã hóa chỉ mục, sắp xếp, lưu trữ và xuất bản CSTT. Để cực tiểu hóa kích thước CSDL luật, chúng tôi chọn thuật toán CRC32 nhằm tránh xung đột giá trị băm cho 4,294,967,296 luật thành viên khác nhau. Do số luật nhận dạng luôn nhỏ hơn số mẫu dữ liệu nên CRC32 là lựa chọn kinh tế, đảm bảo các luật thành viên hoạt động tốt trên CSDL mẫu có hơn 4.2 tỷ chữ ký mã độc.

Trong giai đoạn xử lý, phần mềm được thiết kế bằng ngôn ngữ Object Pascal sử dụng trình biên dịch Delphi XE3 (2012) của Embarcadero Technologies Inc. Gói D2 Anti-virus* 2013 dành cho máy chạy hệ điều hành Windows XP Service Pack 3. Máy tính thử nghiệm sử dụng Intel CPU Core 2 Duo E7200 – 2.53 GHz; 4 GB RAM vật lý; bộ nhớ khả dụng 2.8 GB.

Quá trình thử nghiệm D2 Anti-virus* 2013 (Hình 5) sử dụng CSDL 615,880 mã độc cập nhật ngày 04-08-2013. Mã động cơ 0013.06.01. Tập dữ liệu kiểm tra gồm 8,696 tập thực thi dung lượng 105,330 MB của Windows XP. Chúng tôi thực hiện thử nghiệm 3 lần, ghi nhận số liệu, tính toán, lấy giá trị trung bình. Kết quả tốc độ quét trung bình của D2 đạt 21,651 MB/giây. Kiểm tra 1 mẫu dữ liệu 12.1MB, D2 chỉ cần 0.574977 miligiây. Thời gian kiểm tra 1MB dữ liệu là 0.04747 miligiây (Bảng 3).



Hình 5: Thử nghiệm D2 Anti-Virus 2013*

Bảng 3: Số liệu thực nghiệm D2 Anti-virus 2013*

Lần kiểm tra	Thời gian kiểm tra (giây)	Tốc độ kiểm tra (MB/giây)	Thời gian kiểm tra trung bình (miligiây)	
			1 mẫu	1 MB
1	6	17,555	0.689972	0.05696
2	5	21,066	0.574977	0.04747
3	4	26,332	0.459982	0.03798
Trung bình	5	21,651	0.574977	0.04747

4.2 Tham khảo hệ anti-virus khác

Trong thực nghiệm, chúng tôi gặp khó khăn khi so sánh tốc độ quét của D2 với các hệ AV khác. Do mỗi hệ AV sở hữu một CSDL mẫu có số chữ ký mã độc khác nhau nên không thể so sánh tốc độ hai phần mềm khác nhau về kích thước tập mẫu. Ngoài ra, cách thức phân tích dữ liệu, kỹ thuật so khớp mẫu, nhu cầu tài nguyên, xử lý luồng... của mỗi AV cũng ảnh hưởng tốc độ quét của phần mềm [7].

Để đánh giá kết quả nghiên cứu, chúng tôi khảo sát phần mềm Avira Antivirus với giả định kích thước tập mẫu ảnh hưởng nhiều đến tốc độ so với các yếu tố khác. Avira (Hình 6) là AV miễn phí của Avira GmbH & Co. KG (Germany). Thử nghiệm Avira phiên bản 10.0.0.4052 (mã động cơ 8.02.12.112; CSDL 150,000 mẫu) trong cùng điều kiện với D2 (hệ điều hành, CPU, RAM, tập dữ liệu kiểm tra...), chúng tôi thu được số liệu thực nghiệm qua 3 lần chạy Avira (Bảng 4).

Hình 6: Khảo sát hệ Avira Antivirus

Bảng 4: Khảo sát hệ Avira Antivirus

Lần kiểm tra	Thời gian kiểm tra (giây)	Tốc độ kiểm tra (MB/giây)	Thời gian kiểm tra trung bình (miligiây)	
			1 mẫu	1 MB
1	1027	102.5609	118.1003	9.75030
2	1050	100.3143	120.7452	9.96867
3	1038	101.4740	119.3652	9.854742
Trung bình	1038.333	101.4497	119.4036	9.857907

Kết quả khảo sát cho thấy D2 có số mẫu nhiều hơn Avira 4 lần nhưng thời gian quét của D2 ít hơn Avira 200 lần (số liệu so sánh tương đối, chỉ mang tính tham khảo, không dùng để đánh giá phần mềm Avira Antivirus).

5 KẾT LUẬN VÀ HƯỚNG NGHIÊN CỨU TƯƠNG LAI

Trong bối cảnh gia tăng các cuộc tấn công mạng, mã độc xuất hiện ngày càng nhiều, quét nhanh chữ ký mã độc luôn là bài toán đặt ra cho các hệ AV hiện nay. Để cải tiến tốc độ duyệt quét của các hệ AV dành cho máy tính cá nhân, chúng tôi nghiên cứu cài đặt kỹ thuật nhận dạng mã độc sử dụng cơ chế băm theo chỉ mục trên không gian phân hoạch cho hệ phần mềm D2 Anti-virus* 2013. Không chỉ giải quyết bài toán tăng tốc truy xuất tập mẫu của các AV, kỹ thuật này có thể ứng dụng để lập chỉ mục cho các hệ quản trị CSDL quan hệ kích thước dữ liệu lớn.

Trong thời gian tới, chúng tôi sẽ tiếp tục nghiên cứu phương pháp sử dụng tập luật trong các hệ quản trị CSTT làm tập chỉ mục tìm kiếm khai thác dữ liệu. Kết quả thử nghiệm chứng tỏ kỹ thuật này giúp cải thiện tốc độ cho các hệ duyệt quét mã độc, mở ra hướng giải quyết các bài toán khai thác dữ liệu, tối ưu mô hình quan hệ CSDL và CSTT cho các hệ học ngày nay.

TÀI LIỆU THAM KHẢO

1. Dantong Yu, Aidong Zhang, 2003. ClusterTree: Integration of Cluster Representation and Nearest – Neighbor Search for Large Data Sets with High Dimension. IEEE Transaction on Knowledge and Data Engineering. Vol. 15, No. 3. 1-23.
2. Hoàng Kiếm, Trương Minh Nhật Quang, 2008. Cơ chế máy học chẩn đoán virus máy tính. Tạp chí Tin học và Điều khiển học. Số 1 (2008), Tập 24, Việt Nam. 32-41.
3. Jelena Mirkovic, Peter Reiher, 2004. A taxonomy of DDoS attack and DDoS defense mechanisms. ACM SIGCOMM Computer Communication. Vol. 32 Issue 2. 39-53.
4. Joseph Rabaiotti, 2007. Counter Intrusion Software. PhD. Thesis, Computer Science, Cardiff University. 38-43.
5. Konstantin Rozinov, 2005. An Abstract Efficient Static Analysis of Executables for Detecting Malicious Behaviors. Master of Science Thesis. Brooklyn Polytechnic University. USA.
6. Maurício R. Mediano, Marco A. Casanova, Marcelo Dreux , 1994. V-Trees, A Storage Method for Long Vector Data. Proceedings of the 20th VLDB Conferenc. Santiago - Chile.
7. Ozgun Erdogan, Pei Cao, 2007. Hash-AV: Fast Virus Signature Scanning by Cache-Resident Filters. International Journal of Security and Networks. Volume 2 Issue 1/2. 50-59
8. Peter Szor, 2005. The Art of Computer Virus Research and Defense. Addison Wesley Professional Press (ISBN 0-321-30454-3)
9. Po-Ching Lin, Ying-Dar Lin, Yuan-Cheng Lai, 2011. Hybrid Algorithm of Backward Hashing And Automaton Tracking. IEEE Transactions On Computers. Vol. 60. No. 4. 594-601.
10. Truong Minh Nhat Quang, Hoang Trong Nghia, 2008. A Multi-agent Mechanism in Machine Learning Approach to Anti-virus System. The Proceedings of the 2nd Symposium on Agents and Multi-Agent Systems, KES-AMSTA, Korea. Springer

Lecture Notes in Artificial Intelligence, Vol. 4953, 743-752.

11. Trương Minh Nhật Quang, Hoàng Kiếm, Nguyễn Thanh Thủy, 2008. Ứng dụng Máy học và Hệ chuyên gia trong phân loại và nhận dạng virus máy tính. Tạp chí Công nghệ Thông tin và Truyền thông (ISSN 0866-7039). Số 19,2-2008, Việt Nam. 93-101.
12. US Patent 6763466, July 2004. Fast virus scanning. Inventor – Glover. Assignee Networks Associates Technology.
13. US Patent 6898712, May 2005. Test driver ordering. Inventor - Vignoles et al. Assignee Networks Associates Technology.
14. US Patent 6928555, Aug 2005. Method and apparatus for minimizing file scanning by anti-virus programs. Inventor - Drew. Assignee Networks Associates Technology.
15. US Patent 6952776, Oct 2005. Method and apparatus for increasing virus detection speed using a database. Inventor – Chess. Assignee International Business Machines Corporation.
16. US Patent 7036147, April 2006. System, method and computer program product for eliminating disk read time during virus scanning. Inventor – Hursey. Assignee McAfee.
17. US Patent 7043634, May 2006. Detecting malicious alteration of stored computer files. Inventor - Wolff et al. Assignee McAfee.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

NÂNG CAO ĐỘ CHÍNH XÁC PHÂN LOẠI LỚP ÍT MẪU TỪ TẬP DỮ LIỆU MẤT CÂN BẰNG

Bùi Minh Quân[1], Phạm Xuân Hiền[1] và Huỳnh Xuân Hiệp[1]
[1] *Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Improving prediction of the minority class in an imbalance dataset

Từ khóa:
Học với chi phí nhạy cảm, tập đa lớp, dữ liệu mất cân bằng

Keywords:
Cost-sensitive learning, multi-class, imbalanced data

ABSTRACT

A dataset is called imbalance if it has some classes containing more instances than others. In this case, accurately classifying samples in small classes is very difficult. The higher the imbalanced ratio, the more difficult getting a good solution. Cost-sensitive learning is an effective solution for the imbalanced problem. In this paper, we present a decision system with misclassification cost. The system improves the degree of precision in the minor classes which are interested in imbalanced dataset. The system is based on the study of methods of classifying on the imbalanced dataset by cost-sensitive. This system is applied in medical diagnostic. The experimental results show that the accuracy of the diagnostic system is improved.

TÓM TẮT

Vấn đề mất cân bằng dữ liệu xảy ra khi trong tập dữ liệu có lớp chứa số mẫu nhiều hơn các lớp khác. Phân loại chính xác cho mẫu thuộc lớp nhỏ trong tập mất cân bằng là khó khăn. Khi tỷ lệ mất cân bằng của tập dữ liệu càng cao thì việc phát hiện được mẫu của lớp nhỏ càng khó. Học với chi phí nhạy cảm là giải pháp hiệu quả để giải quyết vấn đề mất cân bằng. Trong bài báo này, chúng tôi trình bày một hệ thống gọi là hệ thống quyết định với chi phí, hệ thống giúp cải thiện khả năng phân loại chính xác của lớp nhỏ trong tập dữ liệu mất cân bằng, lớp dữ liệu rất được quan tâm. Hệ thống được xây dựng dựa vào kết quả nghiên cứu giải pháp phân loại trên dữ liệu mất cân bằng tiếp cận với chi phí nhạy cảm. Hệ thống được áp dụng vào lĩnh vực chẩn đoán y học, kết quả thực nghiệm cho thấy khả năng phát hiện chính xác bệnh nhân của hệ thống chẩn đoán được cải thiện.

1 GIỚI THIỆU

Dữ liệu thu được trong các ứng dụng thực tế thường là các tập dữ liệu mất cân bằng. Tập mất cân bằng thường xuất hiện trong các lĩnh vực như chẩn đoán y tế [24], giám sát hệ thống mạng, phát hiện xâm nhập hệ thống [10]... Thông thường trong những lĩnh vực này lớp cần quan tâm lại có rất ít mẫu (lớp nhỏ) [1][9] so với các lớp khác trong tập dữ liệu. Lớp bệnh nhân có rất ít mẫu so với các lớp khác trong ứng dụng y học, giao dịch tấn công có rất ít mẫu so với các lớp giao dịch khác của hệ thống mạng. Việc chẩn đoán đúng nhãn của mẫu thuộc lớp nhỏ là cần thiết và quan trọng. Nếu mẫu thuộc lớp nhỏ chẩn đoán nhãn sai thì giá phải trả là cao hơn nhiều so với chẩn đoán sai nhãn cho mẫu thuộc lớp lớn [5][11]. Các mẫu thuộc các lớp khác nhau khi chẩn đoán sai thì giá phải trả cũng không giống nhau. Chúng tôi gọi đây là bài toán phân loại (chẩn đoán nhãn) có đặc điểm là lớp có số mẫu nhỏ gắn với chi phí chẩn đoán sai lớn và lớp có số mẫu lớn gắn với chi phí chẩn đoán sai nhỏ.

Tỷ lệ mất cân bằng của tập dữ liệu ảnh hưởng rất lớn đến kết quả gán nhãn của lớp ít mẫu. Tỷ lệ mất cân bằng được xem là tỷ lệ so sánh sự phân phối mẫu của các lớp trong tập dữ liệu với lớp có số mẫu nhỏ nhất [6]. Các giải thuật phân loại truyền thống luôn cố gắng cực đại hóa độ chính xác. Các giải thuật này có xu hướng gán nhãn cho mẫu chưa xác định là thuộc lớp chiếm số mẫu lớn và bỏ qua lớp nhỏ [4]. Nếu dựa vào nguyên tắc số đông gán nhãn cho mẫu trong tập mất cân bằng thì độ chính xác khi phân loại trên tập dữ liệu dễ dàng đạt tới ≈ 99% trong khi độ chính xác của lớp nhỏ là ≈ 0%.

Để cải thiện độ chính xác phân loại trên lớp ít mẫu và giữ được độ chính xác trên toàn tập ở mức chấp nhận, bài báo đề xuất xây dựng hệ thống quyết định với chi phí (Decision System with Misclassification Cost). Hệ thống được xây dựng dựa trên nền tảng nghiên cứu thành tựu của phương pháp học với chi phí nhạy cảm trên tập dữ liệu mất cân bằng [1][7][15][26]. Hệ thống cho phép mở rộng áp dụng vào nhiều lĩnh vực giải quyết vấn đề mất cân bằng của ứng dụng cụ thể.

Nội dung phần 2, chúng tôi sẽ trình bày các nghiên cứu liên quan và lý thuyết xây dựng hệ thống quyết định với chi phí. Tiếp theo trong phần 3, chúng tôi mô tả các chức năng trong hệ thống quyết định với chi phí. Kết quả thực nghiệm sẽ được trình bày trong phần 4, kết luận trong phần 5.

2 CÁC NGHIÊN CỨU LIÊN QUAN

Trong những năm gần đây, học với chi phí đã thu hút được nhiều quan tâm của máy học và cộng đồng khai mỏ dữ liệu [26]. Đã có nhiều nghiên cứu học với chi phí nhạy cảm được thực hiện. Những nghiên cứu học tiếp cận với chi phí chia làm hai dạng. Dạng thứ nhất, chi phí gắn theo mẫu, mỗi mẫu có một chi phí đính kèm [1]. Dạng thứ hai là chi phí theo lớp, mỗi lớp có một chi phí, các mẫu trong cùng lớp có cùng chi phí [4][8][9]. Chi phí gắn với giải thuật học có thể chia làm nhiều loại: chi phí kiểm tra [14], chi phí huấn luyện [24], chi phí phân loại sai [26],... Trong phạm vi bài báo này chúng tôi nghiên cứu chi phí phân loại sai theo lớp.

Tiếp cận với chi phí là giải pháp hiệu quả để giải quyết vấn đề mất cân bằng [15]. Để giải quyết vấn đề mất cân bằng, điều chỉnh tỷ lệ là cách tiếp cận phổ biến của phương pháp học với chi phí. Tuy nhiên, điều chỉnh mất cân bằng theo phương pháp truyền thống chỉ mang lại hiệu quả cao khi áp dụng cho tập hai lớp [26]. Trên tập đa lớp có thể chia làm điều chỉnh trực tiếp (đồng thời) và điều chỉnh gián tiếp. Điều chỉnh đồng thời mất cân bằng trên tập đa lớp chỉ thực hiện khi tìm được bộ chi phí phân loại sai thích hợp giữa các lớp. Nếu không tìm được thì phải điều chỉnh gián tiếp tập đa lớp thông qua các tập con hai lớp [2][26]. Giải thuật học dùng chi phí như là một đại lượng (trọng số) điều chỉnh tỷ lệ mất cân bằng của tập. Trọng số được gán cho mẫu và đưa vào giải thuật học cây quyết định (C4.5) huấn luyện mô hình phân loại [20][21], mô hình sẽ quan tâm đến các mẫu có trọng số cao trong quá trình phân loại [21].

3 PHÂN LOẠI DỮ LIỆU TIẾP CẬN VỚI CHI PHÍ NHẠY CẢM

3.1 Điều chỉnh mất cân bằng tập hai lớp

Trong tập dữ liệu hai lớp, các giải thuật điều chỉnh mất cân bằng với chi phí thì mẫu được phân loại thuộc về lớp có tổng chi phí tổn thất là thấp nhất [8][9].

Ví dụ 1: với tập dữ liệu hai lớp, p là xác suất phân loại mẫu thuộc lớp 1 và 1-p là xác suất mẫu thuộc lớp 2. Các giá trị cm_{ij} của ma trận chi phí là chi phí phân loại sai mẫu thuộc lớp i vào lớp j (i≠j), $i, j \in \{1..2\}$, $cm_{ij}=0$ (với i=j) là chi phí phân loại đúng.

$$\mathrm{cm} = \begin{bmatrix} 0 & cm_{12} \\ cm_{21} & 0 \end{bmatrix}$$

Mẫu được gán vào lớp 1 khi và chỉ khi tổng chi phí phân loại mẫu vào lớp 1 nhỏ hơn tổng chi phí phân loại mẫu vào lớp 2 [26], công thức tính tổng chi phí như sau:

$$p \times cm_{11} + (1-p) \times \mathrm{cm}_{21} \le \mathrm{p} \times \mathrm{cm}_{12} + (1-\mathrm{p}) \times \mathrm{cm}_{22} \quad (1)$$

Từ công thức (1) cho thấy việc gán nhãn mẫu phụ thuộc vào hai đại lượng: xác suất và chi phí. Khi chi phí thay đổi nhãn gán cho mẫu sẽ thay đổi. Chứng tỏ rằng việc gán nhãn cho mẫu nhạy cảm với chi phí. Chi phí của lớp là đại lượng điều chỉnh mất cân bằng các lớp trong tập. Chi phí cao được gán cho các lớp ít mẫu, chi phí thấp được gán cho các lớp nhiều mẫu. Chi phí giúp cân bằng so với số mẫu. Trọng số gán cho mẫu trong tập huấn luyện được tính từ chi phí [21][26]. Trong quá trình phân loại, mô hình sẽ quan tâm đến các mẫu có trọng số cao, nghĩa là quan tâm đến các lớp ít mẫu. Nếu gọi cm_i là chi phí của lớp thứ i thì cm_i được tính như sau:

$$cm_i = \sum_{j=1}^{k} cm_{ij} \quad (2)$$

Với k=2 (k là số lớp của tập dữ liệu), thì $cm_1=cm_{12}$, $cm_2=cm_{21}$. Gọi w_i là trọng số gán mẫu thuộc lớp thứ i, w_i được tính như sau:

$$w_i = \frac{(num \times cm_i)}{\sum_{j=1}^{k}(num_j \times cm_j)} \quad (3)$$

Với num: số mẫu tập huấn luyện, num_j: số mẫu của lớp thứ j trong tập huấn luyện.

3.2 Điều chỉnh mất cân bằng trên tập đa lớp

Theo kết quả của nghiên cứu điều chỉnh mất cân bằng trên tập đa lớp đã chỉ rõ: phương pháp điều chỉnh truyền thống chỉ cho kết quả tốt trên tập hai lớp, trên tập đa lớp độ chính xác phân loại chưa tốt [26]. Nguyên nhân là chưa tìm ra được bộ trọng số thích hợp cho phép điều chỉnh mất cân bằng trên tập đa lớp.

Nghiên cứu đã chỉ ra rằng, điều chỉnh tỷ lệ trên tập dữ liệu k lớp chỉ thực hiện trực tiếp khi tìm được bộ trọng số $w[w_1,w_2,..,w_k]$ thích hợp, với w_i là trọng số của lớp thứ i [26]. Giả sử cần tìm bộ trọng số w cho tập dữ liệu k lớp $i, j \in \{1..k\}$, ma trận chi phí cm như sau:

$$cm=\begin{bmatrix} 0 & cm_{12} & cm_{13} & .... & cm_{1k} \\ cm_{21} & 0 & cm_{23} & .... & cm_{2k} \\ .... & .... & .... & .... & .... \\ cm_{k1} & cm_{k2} & cm_{k3} & .... & 0 \end{bmatrix}$$

Giải phương trình hệ số (4) với k biến, nếu tìm được nghiệm của phương trình (4) $w_=[w_1,w_2,..,w_k]^T$ thì các lớp có thể điều chỉnh mất cân bằng đồng thời [26]. Bộ nghiệm được xem dùng như là bộ trọng số điều chỉnh mất cân bằng đồng thời các lớp khi chi phí phân loại sai mẫu của các lớp là bằng nhau (w=w_). Trong thực tế chi phí phân loại sai mẫu của các lớp là không bằng nhau. Bộ nghiệm w_ tìm được từ phương trình (4) được xem là bộ chi phí phân loại sai của lớp (cm=w_). Bộ trọng số của lớp $w=[w_1,w_2,..,w_k]$ được tính từ công thức (3) (tính từ chi phí của lớp, số mẫu của lớp và số mẫu tập huấn luyện) [26].

$$\begin{cases} w_1 \times cm_{21} - w_2 \times cm_{12} + w_3 \times 0 + ... + w_k \times 0 = 0 \\ w_1 \times cm_{31} + w_2 \times 0 - w_3 \times cm_{13} + ... + w_k \times 0 = 0 \\ ... \quad ... \quad ... \quad ... \quad ... \quad ... \quad ... \quad ... = 0 \\ w_1 \times cm_{k1} + w_2 \times 0 + w_3 \times 0 + ... - w_k \times cm_{1k} = 0 \\ w_1 \times 0 + w_2 \times cm_{32} - w_3 \times cm_{23} + ... + w_k \times 0 = 0 \\ ... \quad ... \quad ... \quad ... \quad ... \quad ... \quad ... \quad ... = 0 \\ w_1 \times 0 + w_2 \times cm_{k2} + w_3 \times 0 + ... - w_k \times cm_{2k} = 0 \\ ... \quad ... \quad ... \quad ... \quad ... \quad ... \quad ... \quad ... = 0 \\ ... \quad ... \quad ... \quad ... \quad ... \quad ... \quad ... \quad ... = 0 \\ w_1 \times 0 + w_2 \times 0 + w_3 \times 0 + ... - w_k \times cm_{k-1,k} = 0 \end{cases} \quad (4)$$

Bộ nghiệm chỉ tìm được khi và chỉ khi hạng của ma trận hệ số M được trình bày theo công thức (5) là nhỏ hơn k (Rank(M)<k) [26]. Điều này cũng ít khi xảy ra, vì hạng của một ma trận $\frac{k(k-1)}{2} \times k$ với (k>2) hầu như luôn bằng k.

$$\begin{bmatrix} cm_{21} & -cm_{12} & 0 & ... & 0 \\ cm_{31} & 0 & -cm_{13} & ... & 0 \\ ... & ... & ... & ... & 0 \\ cm_{k1} & 0 & 0 & ... & -cm_{1k} \\ 0 & cm_{32} & -cm_{23} & ... & 0 \\ ... & ... & ... & ... & 0 \\ 0 & cm_{k2} & 0 & & -cm_{2k} \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ 0 & ... & 0 & ... & -cm_{k-1,k} \end{bmatrix} \quad (5)$$

Khi hạng của ma trận hệ số M là k (Rank(M) = k), có nghĩa là không có bộ nghiệm nào được tìm thấy, tức không tìm được bộ trọng số thích hợp áp dụng điều chỉnh trực tiếp tỷ lệ đồng thời giữa các lớp [26]. Như vậy thay vì điều chỉnh trực tiếp trên tập đa lớp, chúng ta đưa về thành điều chỉnh gián tiếp [2]. Bằng cách điều chỉnh mất cân bằng trên các tập hai lớp và ma trận chi phí của hai lớp theo phương pháp truyền thống [4][7]. Các tập hai lớp là các cặp lớp được tách ra từ tập đa lớp, số tập hai lớp bằng tổ hợp chập 2 của k lớp. Ma trận chi phí của hai lớp được tách từ ma trận chi phí tập đa lớp. Trọng số được tính từ ma trận chi phí. Trọng số được gán cho mẫu trong tập huấn luyện. Giải thuật học dựng mô hình phân loại từ tập mẫu đã gán trọng số. Kết quả phân loại mẫu là kết quả được bình chọn từ tập mô hình hai lớp [26].

Tóm tắt điều chỉnh mất cân bằng trên tập đa lớp tiếp cận với chi phí phân loại sai [26].

Đầu vào: Tập huấn luyện D (với k lớp), ma trận chi phí, giải thuật học G(C4.5/J48)

1. Tạo ma trận hệ số M từ ma trận chi phí Matrix(cm)
2. Kiểm tra hạng của M nếu Rank(M) < k
3. Tìm ra bộ vector trọng số $w=[w_1,w_2,..,w_k]$, gán trọng số cho mẫu trong tập D, kết quả thu được tập D*. Dùng G dựng nên mô hình H từ tập dữ liệu D^* ($H=G(D^*)$).
4. Ngược lại

5. for i = 1 to k-1 do
6. for j = i + 1 to k do
7. Từ tập D lấy ra tập D_{ij}, D_{ij} là tập dữ liệu có 2 lớp chỉ chứa các mẫu thuộc lớp i và lớp j. $cm_{ij} = \begin{bmatrix} cm_{ii} & cm_{ij} \\ cm_{ji} & cm_{jj} \end{bmatrix}$ là ma trận chi phí của lớp i và j tách ra từ ma trận cm.
8. $h_{ij} \leftarrow$ Traditional_rescaling(D_{ij},cm_{ij},G)

% áp dụng phương pháp điều chỉnh tỷ lệ truyền thống trên tập hai lớp D_{ij}.

9. End for
10. End for
11. H(x)$\leftarrow$

$$\arg\max_{y\in\{1,..,k\}} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} I(h_{ij}(x) = y)$$

% kết quả phân loại của mẫu x là sự bình chọn nhãn từ tập mô hình phân loại h_{ij}

12. end if

Đầu ra: mô hình phân loại với chi phí.

4 HỆ THỐNG QUYẾT ĐỊNH VỚI CHI PHÍ

4.1 Xây dựng hệ thống quyết định với chi phí

Dựa trên kết quả nghiên cứu của phương pháp điều chỉnh tỷ lệ mất cân bằng trên tập đa lớp tiếp cận với chi phí. Trong phần này chúng tôi đưa ra định nghĩa cho hệ thống quyết định truyền thống (DS– Decision System) có 5 thành phần như sau:

Định nghĩa 1: S= (U, C, d, V, f)

Trong đó, U là tập không gian mẫu, không rỗng. C là tập các thuộc tính điều kiện và d là thuộc tính quyết định. Q là tập các thuộc tính, Q = U $\cup$ d. $V = U_{q\in Q} V_q$ với V_q là miền giá trị của thuộc tính q. f: U $\times$ Q $\rightarrow$ V là tổng chức năng quyết định (chức năng thông tin). Như vậy f (x,q) $\in V_q$ với q $\in$ Q, x $\in$ U [12].

Dựa trên định nghĩa 1, mở rộng định nghĩa cho hệ thống quyết định với chi phí có 7 thành phần sau:

S= (U, C, d, V, f, cm, n)

Các thành phần U, C, d, V, f tương tự như định nghĩa 1. Với k=$|V_d|$ là miền giá trị của thuộc tính quyết định (số lớp của tập dữ liệu). Khi đó n $\in$ $R^+ \cup \{0\}$ là giá trị lớn nhất có thể gán các phần tử trong ma trận chi phí cm. mc: k $\times$ k $\rightarrow$ $R^+ \cup \{0\}$, với $i, j \in \{1..k\}$, 0<=cm_{ij}<=n, cm_{ii} =0 chi phí gán nhãn đúng mẫu, cm_{ij} (nếu i $\neq$ j) là chi phí phải trả khi gán nhãn sai mẫu (mẫu thuộc lớp i được gán nhãn vào lớp j).

Đơn vị đo lường chi phí: thực tế để xây dựng ma trận chi phí phải dựa vào từng lĩnh vực chuyên môn. Đơn vị của chi phí có thể là tiền, thời gian tính toán, thời gian khắc phục hậu quả, sức khỏe bệnh nhân...

Ví dụ 1: trong tập dữ liệu ann [3] ghi nhận kiểm tra mắc bệnh suy giảm tuyến giáp của 7200 bệnh nhân. Mỗi bệnh nhân được kiểm tra tuyến giáp và chẩn đoán vào một trong ba nhóm: lớp 1 là có bệnh, có 166 mẫu; lớp 2 có tuyến giáp dưới mức bình thường, có 368 mẫu và lớp 3 là bình thường, có 6666 mẫu. Tập có phân phối mẫu các lớp có tỷ lệ [1:2.2:40.1], tập có tỷ lệ mất cân bằng cao [6] và ma trận chi phí của tập có dạng như sau:

$$cm = \begin{bmatrix} cm_{11} & cm_{12} & cm_{13} \\ cm_{21} & cm_{22} & cm_{23} \\ cm_{31} & cm_{32} & cm_{33} \end{bmatrix}$$

Để xây dựng được ma trận chi phí, cần xác định các giá trị của ma trận, trong đó cm_{11} là chi phí khắc phục hậu quả một người bệnh được chẩn đoán là bệnh (chi phí bằng 0); cm_{12} là chi phí khắc phục hậu quả một người bệnh được chẩn đoán là dưới bình thường; cm_{13} là chi phí khắc phục hậu quả một người bệnh được chẩn đoán là bình thường; mở rộng tính toán các giá trị còn lại. Trong thực tế việc phân loại sai mẫu vào các lớp khác nhau có chi phí khác nhau.

Ma trận chi phí được xây dựng từ việc tính toán giá phải trả khi phân loại sai mẫu của một lớp vào các lớp còn lại. Thông thường việc ước lượng này là mất nhiều công sức tính toán. Trong một số lĩnh vực ma trận chi phí được xây dựng từ tri thức chuyên gia [26].

4.2 Chỉ số đánh giá

Trên tập dữ liệu đa lớp mất cân bằng số mẫu lớp nhỏ là rất ít, tỷ lệ phân loại chính xác trên lớp ít mẫu là rất được quan tâm. Trong quá trình thực nghiệm, kết quả phân loại mẫu được trình bày tại ma trận confusion matrix [26]. Kết quả được so sánh trên ba chỉ số đo rất được quan tâm trong các ứng dụng chẩn đoán y học là Accuracy (độ chính

xác toàn tập) và $Recall_i$ (độ bao phủ lớp thứ i), $Precision_i$ độ chính xác lớp thứ i.

Bảng 1: Confusion matrix trên tập đa lớp

Lớp đúng	Lớp dự đoán			
	Lớp 1	**Lớp 2**	**…**	**Lớp k**
Lớp 1	N_{11}	N_{12}	…	N_{1k}
Lớp 2	N_{21}	N_{22}	…	N_{2k}
…	…	…	…	…
Lớp k	N_{k1}	N_{k2}	…	N_{kk}

N_{ii} số lượng mẫu thuộc lớp i phân đúng vào lớp i. N_{ij} số lượng mẫu thuộc lớp i phân loại sai vào lớp j (với $i \neq j$).

Accuracy(Acc): độ chính xác phân loại toàn tập dữ liệu là xác suất tính trên số mẫu phân loại đúng trên tổng số mẫu phân loại của tập.

$Recall_i$ (Re_i): độ bao phủ, khả năng phát hiện chính xác một mẫu thuộc lớp thứ i, là xác suất phát hiện chính xác mẫu lớp i trên tổng số mẫu phân loại thuộc lớp i.

$Precision_i$ (Pr_i): độ chính xác của lớp thứ i, là xác suất phân loại chính xác mẫu lớp i trên tổng số mẫu được phân loại thuộc về lớp i.

$$Acc = \frac{\sum_{i=1}^{|k|} N_{ii}}{\sum_{i=1}^{k}\sum_{j=1}^{k} N_{ij}} \qquad \mathrm{Re}_i = \frac{N_{ii}}{\sum_{j=1}^{|k|} N_{ij}}$$

$$\mathrm{Pr}_i = \frac{N_{ii}}{\sum_{j=1}^{|k|} N_{ji}}$$

5 XÂY DỰNG HỆ THỐNG QUYẾT ĐỊNH VỚI CHI PHÍ

Chúng tôi đã xây dựng hệ thống quyết định với chi phí dựa trên phương pháp phân loại dữ liệu mất cân bằng tiếp cận với chi phí. Hệ thống được áp dụng vào chẩn đoán bệnh suy giảm tuyến giáp ở người (ann tập dữ liệu có tỷ lệ mất cân bằng >40) [6]. Tập dữ liệu được tiến hành với hai loạt thực nghiệm điều chỉnh trực tiếp và gián tiếp. Thông qua kết quả thực nghiệm kiểm chứng vai trò của chi phí trong việc cải thiện tỷ lệ phân loại chính xác trên lớp ít mẫu.

Điều kiện cần để thực thi hệ thống là tập huấn luyện, ma trận chi phí và tập kiểm thử. Vì vậy tập dữ liệu được tách ra thành hai tập con: tập huấn luyện và tập kiểm thử theo tỷ lệ 1:1. Giá trị U, C, d, V, f được xác định từ tập huấn luyện, đặt n=10.

Khởi tạo ma trận chi phí cm cho tập dữ liệu ann, chi phí chỉ mang tính chất chất thực nghiệm. Giá trị ma trận chi phí cm thỏa 3 ràng buộc. Thứ nhất, ma trận có ít nhất một giá trị 1 và giá trị của ma trận được sinh ra ngẫu nhiên và phải nhỏ hơn hoặc bằng n. Thứ hai, chi phí phân loại sai của lớp nhỏ nhất vào lớp lớn nhất là lớn nhất. Thứ ba, chi phí phân loại sai của lớp lớn nhất vào lớp nhỏ nhất là nhỏ nhất.

Chúng tôi xây dựng bộ công cụ của hệ thống quyết định với chi phí có các chức năng sau:

5.1 Sinh ma trận chi phí điều chỉnh trực tiếp mất cân bằng trên tập đa lớp

Đặt hằng n=10, khởi tạo ma trận cm(k, k) với k là số lớp của tập dữ liệu, $i, j \in \{1..k\}$, cm[i,j] $\in R^+ \cup \{0\}$, 0<=cm[i,j] <=n (cm_{ij} được sinh ngẫu nhiên). Khởi tạo vector cv với k phần tử, $i \in \{1..k\}$, cv[i] $\in R^+ \cup \{0\}$, 0<=cv[i]<=n (cv[i] được sinh ngẫu nhiên).

Kiểm tra điều kiện của ma trận chi phí cm thỏa ba điều kiện đã trình bày ở trên. Kiểm tra cv có phải là nghiệm của ma trận hệ số từ ma trận chi phí cm theo công thức (4), hạng của ma trận hệ số theo công thức (5) là nhỏ hơn k. Nếu đúng thì vector cv được dùng tính trọng số w, trọng số w dùng điều chỉnh mất cân bằng trực tiếp trên tập đa lớp. Chi phí này gọi là chi phí nhất quán. Giá trị vector cv được kiểm tra thỏa 3 ràng buộc như ma trận chi phí trước khi gán cho lớp [26].

5.2 Sinh ma trận chi phí điều chỉnh gián tiếp mất cân bằng trên tập đa lớp

Đặt hằng n=10, khởi tạo ma trận cm(k,k) với k là số lớp của tập dữ liệu, $i, j \in \{1..k\}$, các giá trị của ma trận được sinh ra ngẫu nhiên tương tự như ma trận chi phí điều chỉnh trực tiếp (0<=cm[i,j] <=n). Kiểm tra điều kiện của ma trận chi phí cm thỏa ba điều kiện ràng buộc được trình bày như trên, ngoài ra hạng (Rank) của ma trận hệ số theo công thức (5) là không nhỏ hơn k. Ma trận chi phí này được dùng điều chỉnh gián tiếp trên tập đa lớp, chi phí này gọi là chi phí không nhất quán [24][26].

5.3 Dựng mô hình phân loại với chi phí

Trọng số w_i gán cho mẫu lớp i được tính từ chi phí của lớp. Khi điều chỉnh trực tiếp chi phí lớp là nghiệm của phương trình (4). Khi điều chỉnh gián tiếp chi phí tập hai lớp được tính theo công thức (2). Áp dụng công thức (3) tính trọng số w_i của lớp

thứ i. Trọng số lớp dùng gán cho các mẫu thuộc lớp trong tập huấn luyện. Tập huấn luyện đã gán trọng số đưa vào giải thuật học C4.5 (trên java C4.5 là J4.8). Kết quả thu được là mô hình phân loại với chi phí.

5.4 Chức năng phân loại

Tập kiểm thử được đưa vào mô hình phân loại với chi phí. Kết quả phân loại được trình bày qua ma trận confusion matrix và thống kê các chi số đo Acc, Re_i, Pr_i (i là lớp được quan tâm).

6 KẾT QUẢ THỰC NGHIỆM

Tập dữ liệu ann thuộc lĩnh vực chẩn đoán tuyến giáp, tập được chọn từ UCI (University of California, Irvine) của Blake *et at*., 1998 [3]. Tập dữ liệu ann có 7200 mẫu phân phối vào 3 lớp, được tài trợ bởi Randolf Werner. Lớp 1 là có bệnh, 166 mẫu; lớp 2 tuyến giáp dưới mức bình thường, 368 mẫu và lớp 3 là bình thường, 6666 mẫu [166;368;6666]. Tập dữ liệu có 21 thuộc tính (15 thuộc tính kiểu nhị phân, 6 thuộc tính kiểu liên tục).

Tập dữ liệu ann được tách ra là 2 tập con: tập huấn luyện (tập train) [83;184;3333] và tập kiểm thử (tập test) [83;184;3333]. Lớp được quan tâm đến tỷ lệ phân loại chính xác là lớp 1 (lớp bệnh).

Thực nghiệm 1: thực hiện dựng mô hình với giải thuật J4.8 và tập huấn luyện với các mẫu không gán trọng số [25]. Dùng tập test kiểm thử, kết quả ghi nhận trong thực nghiệm như sau:

Acc = 99.47 %, Re_1 = 94.00 %, Pr_1 = 95.12%

Bảng 2: Confusion matrix trên TN1

Lớp đúng	Lớp dự đoán		
	1	**2**	**3**
1	78	0	5
2	0	175	9
3	4	1	3328

Thực nghiệm 2 (TN2): Kết quả thực nghiệm khi điều chỉnh trực tiếp với chi phí nhất quán. Khởi tạo ma trận chi phí. Bộ nghiệm tìm được là cv={1.0000; 0.8686; 0.5992}. Như vậy chi phí lớp 1 là 1.000, chi phí lớp 2 là 0.8686, chi phí lớp 3 là 0.5992. Chi phí lớp được dùng tính trọng số lớp, trọng số gán cho mẫu trong tập train. Dựng mô hình từ tập train đã được gán trọng số. Mô hình được dùng kiểm thử với tập test. Kết quả ghi nhận trong thực nghiệm như sau:

Acc = 99.61 %, Re_1 = 98.80 %, Pr_1 = 96.47%

Bảng 3: Confusion matrix trên TN2

Lớp đúng	Lớp dự đoán		
	1	**2**	**3**
1	82	0	1
2	0	177	7
3	3	3	3327

Bảng 4: Thống kê sự thay đổi phân phối mẫu trên tập (điều chỉnh trực tiếp)

Lớp dữ liệu	chi phí nhất quán		không chi phí	
	Số mẫu được dự đoán	***Số mẫu đúng lớp***	***Số mẫu được dự đoán***	***Số mẫu đúng lớp***
1	85	82	82	78
2	180	177	176	175
3	3335	3327	3342	3328

Thực nghiệm 3: Kết quả thực nghiệm khi điều chỉnh gián tiếp với chi phí không nhất quán. Khởi sinh ma trận chi phí không nhất quán, tập huấn luyện được tách ra thành tổ hợp chập 2 của k các tập hai lớp, ma trận chi phí của các tập hai lớp được tách ta từ ma trận sinh ra ban đầu. Sau đó, xây dựng các mô hình từ các tập hai lớp và ma trận chi phí tương ứng. Sử dụng tập kiểm tra để phân loại dữ liệu dựa vào việc bình chọn từ tập các mô hình. Kết quả thực nghiệm như sau:

Acc = 99.56 %, Re_1 = 96.39 %, Pr_1 = 95.23%

Bảng 5 : Confusion matrix trên TN3

Lớp đúng	Lớp dự đoán		
	1	**2**	**3**
1	80	0	3
2	0	176	8
3	4	1	3328

Bảng 6: Thống kê sự thay đổi phân phối mẫu trên tập (điều chỉnh gián tiếp)

Lớp dữ liệu	chi phí điều chỉnh gián tiếp		không chi phí	
	Số mẫu được dự đoán	***Số mẫu đúng lớp***	***Số mẫu được dự đoán***	***Số mẫu đúng lớp***
1	84	80	82	78
2	177	176	176	175
3	3339	3328	3342	3328

Phân tích kết quả thực nghiệm, với TN1 số mẫu được phân loại vào lớp 1 là 82 trong đó có 78 mẫu là chính xác thuộc về lớp 1 (78/82). Với TN2 số mẫu được phân loại vào lớp 1 là 85 trong đó có 82 mẫu là chính xác thuộc về lớp 1 (82/85). Với TN3

số mẫu được phân loại vào lớp 1 là 84 trong đó có 80 mẫu là chính xác thuộc về lớp 1 (80/84). Kết quả TN2 và TN3 đều nâng cao số lượng bệnh nhân được chẩn đoán đúng bệnh điều này mang lại cho bệnh nhân cơ hội chăm sóc, cơ hội sống nhờ phát hiện kịp thời, đúng thời điểm. Việc chẩn đoán sai người có bệnh vào các nhóm khác phải trả giá rất cao: sức khỏe, sự sống của bệnh nhân lý do là không phát hiện kịp thời, bỏ qua thời điểm vàng điều trị bệnh.

Từ kết quả thực nghiệm cho thấy đã có sự phân phối mẫu lại giữa các lớp; số lượng mẫu được phân loại có xu hướng di chuyển vào lớp được gán trọng số cao. Sự dịch chuyển này làm nâng cao số lượng mẫu phân loại vào lớp ít mẫu và số mẫu được phân loại đúng trên lớp ít mẫu.

Kết quả phân loại mẫu phụ thuộc nhiều vào giá trị ma trận chi phí. Chúng tôi tiến hành thực nghiệm 10 lần (TN2 và TN3), kết quả được trình bày dưới dạng: trung bình ± độ lệch chuẩn.

Bảng 7: Thống kê trên 3 dạng thực nghiệm

	chi phí nhất quán	chi phí không nhất quán	không chi phí
Acc	0.9962 ± 0.0004	0.9953 ± 0.0003	0.9947
Re (lớp1)	0.9856 ± 0.0069	0.9675 ±0.0108	0.9400
Pr (lớp1)	0.9635 ± 0.0037	0.9525 ±0.0005	0.9512

Đối với các tập dữ liệu mất cân bằng cao, cả chi phí nhất quán và chi phí không nhất quán đều cải thiện được tỷ lệ phân loại chính xác mẫu lớp nhỏ của tập dữ liệu trên hai độ đo Recall và Precision. Độ chính xác phân loại ổn định và không làm giảm tỷ lệ phân loại chính xác trên toàn tập (độ đo Accuracy). Thông qua kết quả thực nghiệm cho thấy được vai trò của chi phí trong quá trình phân loại dữ liệu. Việc gán nhãn cho mẫu nhạy cảm với chi phí, chi phí đã làm thay đổi sự phân phối mẫu giữa các lớp. Sự thay đổi có xu hướng di chuyển về lớp có chi phí phân loại sai lớn. Chứng tỏ việc gán nhãn cho mẫu không chỉ phụ thuộc vào nguyên tắc số đông mà còn phụ thuộc vào chi phí.

7 KẾT LUẬN

Bài báo đã trình bày các bước xây dựng hệ thống quyết định với chi phí phân loại sai. Hệ thống cải thiện được tỷ lệ phân loại chính xác trên lớp ít mẫu trong tập mất cân bằng. Nếu áp dụng hệ thống vào ứng dụng chẩn đoán y học sẽ nâng cao hiệu quả chẩn đoán, nếu áp dụng vào lĩnh vực phát hiện xâm nhập, tấn công sẽ năng cao hiệu quả giám sát hệ thống. Tuy nhiên, chúng tôi chưa đề xuất được một bộ tiêu chí xây dựng ma trận chi phí cho tập dữ liệu mất cân bằng, việc xây dựng ma trận chủ yếu dựa vào tri thức chuyên gia.

TÀI LIỆU THAM KHẢO

1. Abe, N., Zadrozny, B., Langford, J. (2004), An iterative method for multi-class cost-sensitive learning, Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, pp. 3–11.
2. Allwein, E. L., Schapire, R. E., Singer, Y. (2000), Reducing multiclass to binary: A unifying approach for margin classifiers, Journal of Machine Learning Research 1, pp. 113–141.
3. Blake, C., Keogh, E., Merz, C. J. (1998), UCI repository of machine learning databases, [http://www.ics.uci.edu/~mlearn/MLRepository.html], Department of Information and Computer Science, University of California, Irvine, CA.
4. Breiman, L., Friedman, J. H., Olsen, R. A., Stone, C. J. (1984), Classification and Regression Trees. Wadsworth, Belmont, CA.
5. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002), SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, pp. 321–357.
6. Ding, Z. (2011). Diversified Ensemble Classifier for Highly imbalanced Data Learning and their application in Bioinformatics, Ph. D thesis, College of Arts and science, Department of Computer Science, Georgia State University,2011. Http://digitalarchive.gsu.edu/cs_diss/60
7. Domingos, P. (1999), MetaCost: A general method for making classifiers costsensitive, Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, pp. 155–164.
8. Drummond, C., Holte, R. C. (2003), C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling, Working Notes of the ICML'03 Workshop on Learning from Imbalanced Data Sets, Washington, DC.

9. Elkan, C. (2001), The foundations of cost-senstive learning, Proceedings of the 17th International Joint Conference on Artificial Intelligence. Seattle, WA, pp. 973–978.
10. Engen Vegard. 2010. Machine Learning for Network Based Intrusion Detection. Ph. D thesis, Bounemouth University, 2010.
11. Hido, S. and Kashima, H. 2008. Roughly Balanced Bagging for Imbalanced Data. "In Proceedings of SIAM Conference on Data Mining (SDM2008), Atlanta, Georgia, USA, April, 2008.
12. Hong Zhao, Fan Min, William Zhu 2012. Minimal cost feature selection of data with normal distribution measurement errors. Lab of Granular Computing, Zhangzhou Normal University, Zhangzhou 363000, China.
13. Jeffrey P. Bradford., Clayton Kunz., Ron Kohavi., Clifford Brunk., Carla E. Brodley. (1998), Pruning Decision Trees with Misclassification Costs. ECML-98, pp.131-136.
14. Ling, C. X., Yang, Q.,Wang, J., Zhang, S. (2004), Decision trees with minimal costs, Proceedings of the 21st International Conference on Machine Learning. Banff, Canada, pp. 69–76.
15. Liu, X.-Y., Zhou, Z.-H. (2006), The influence of class imbalance on cost-sensitive learning: An empirical study, Proceedings of the 6th IEEE International Conference on Data Mining. Hong Kong, China, pp. 970–974.
16. Lozano, A. C., Abe, N., 2008. Multi-class cost-sensitive boosting with p-norm loss functions, Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, NV, pp. 506–514.
17. Maloof, M. A., 2003. Learning when data sets are imbalanced and when costs are unequal and unknown, Working Notes of the ICML'03Workshop on Learning from Imbalanced Data Sets. Washington, DC.
18. Margineantu, D. (2001), Methods for cost-sensitive learning. Ph.D. thesis, department of Computer Science, Oregon State University, Corvallis, OR.
19. Provost, F, Domingos, P. (2003), Tree induction fof probability-base ranking, Machine Learning 52 (3), 199–215.
20. Quinlan, J. R. (1993), C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, California.
21. Ting, K. M. (2002), An instance-weighting method to induce cost-sensitive trees. IEEE Transactions on Knowledge and Data Engineering 14 (3), 659–665.
22. Turney, P. D. (2000), Types of cost in inductive concept learning, Proceedings of the ICML'2000 Workshop on Cost-Sensitive Learning. Stanford, CA, pp. 15–21.
23. Witten, I. H., Frank, E. (2011), Data Mining: Practical Machine Learning Tools and Techniques, Third Edition. Morgan Kaufmann Publishers. www.mkp.com. ISBN: 978-0-12-374856-0.
24. Yang, Y. and Ma, G. 2010. Ensemble-based Active Learning for Classification Problem. J. Biomedical and Engineering, 2010, 3, pp. 1021- 1028. Published online in SciRes. Http:/www. Scrip.org/journal/jbise.
25. Zadrozny, B., Langford, J., Abe, N. (2002), A simple method for cost-sensitive learning. Tech. rep., IBM.
26. Zhou, Z.-H., Liu, X.-Y. (2006a), On multi-class cost-sensitive learning, Proceeding of the 21st National Conference on Artificial Intelligence. Boston, WA, pp. 567–572.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

SO SÁNH CÁC MÔ HÌNH DỰ BÁO LƯỢNG MƯA CHO THÀNH PHỐ CẦN THƠ

Đỗ Thanh Nghị[1], Phạm Nguyên Khang[1], Nguyễn Nhị Gia Vinh[2] và Văn Phạm Đăng Trí[3]

[1] *Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ*
[2] *Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ*
[3] *Khoa Môi trường& Tài nguyên Thiên nhiên, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
A comparision of rainfall forecast models for Can Tho city - Vietnam

Từ khóa:
Dự báo lượng mưa, hồi qui tuyến tính, k láng giềng, cây quyết định, bagging, rừng ngẫu nhiên, máy học véc-tơ hỗ trợ

Keywords:
Rainfall forecast, linear regression, k nearest neighbors, decision trees, bagging, random forests, support vector machines

ABSTRACT

In recent years, climate change is one of the environmental problems that needs to be studied in the Mekong Delta of Vietnam, especially those in conjunction with temperature and rainfall. As temperature and rainfall changes directly affect agriculture and aquaculture activities - driving factors of the delta's development, the raising question is if such changes could be forecasted with acceptable level of uncertainties. This paper presents algorithms and models of adjusting the forecasted rainfall data obtained from climate data of the SEA-START. A comparison of these forecast models is conducted by forecast error analysis. A case study is experimented by using rainfall data in Can Tho city - Vietnam. The results show that the linear regression model has the greatest forecast error while the non-linear forecast models give better results. The diversity of these forecast models can be applied to solve environmental problems in practice.

TÓM TẮT

Trong những năm gần đây, biến đổi khí hậu là một trong những vấn đề môi trường cần được nghiên cứu ở vùng đồng bằng sông Cửu Long - Việt Nam, đặc biệt là những vấn đề liên quan đến các yếu tố nhiệt độ và lượng mưa. Do sự thay đổi nhiệt độ và lượng mưa ảnh hưởng trực tiếp đến các hoạt động nông nghiệp và nuôi trồng thủy sản - những yếu tố chính dẫn đến sự phát triển của vùng đồng bằng Sông Cửu Long, câu hỏi được đặt ra là liệu những thay đổi về nhiệt độ và lượng mưa có thể được dự báo với độ không chắc chắn ở mức có thể chấp nhận được hay không. Bài báo này trình bày các giải thuật và mô hình dự báo lượng mưa từ nguồn dữ liệu khí hậu của SEA-START. Các mô hình dự báo này được so sánh với nhau bằng phương pháp phân tích lỗi dự báo. Các kết quả trong bài báo này cho thấy mô hình hồi qui tuyến tính có lỗi dự báo cao nhất trong khi các mô hình dự báo phi tuyến cho kết quả dự báo tốt hơn. Tính đa dạng của những mô hình dự báo này có thể được ứng dụng để giải các bài toán môi trường trong thực tiễn.

1 GIỚI THIỆU

Hệ thống khí hậu trái đất bao gồm bốn thành phần: khí quyển, đại dương, khu vực đóng băng và sinh quyển. Các tiến trình vật lý như bức xạ, tuần hoàn và mưa phản ứng với các tiến trình sinh học như tiến trình hấp thu carbon do trồng cây, các biến đổi hóa học để hình thành nên hệ thống khí hậu

biến đổi phức tạp (McKuffie *et al*., 2005). Những biến đổi phức tạp này tác động mạnh mẽ đến sản xuất nông nghiệp ở các nước trên thế giới, đặc biệt là các nước ở vùng nhiệt đới. Ở nước ta, các ảnh hưởng của biến đổi khí hậu đối với nguồn tài nguyên nước lên các lĩnh vực nông nghiệp và thủy sản là mối quan tâm hàng đầu của các nhà nghiên cứu thủy văn học. Theo báo cáo của (Bộ Tài nguyên và Môi trường, 2011) thì vùng đồng bằng sông Cửu Long (ĐBSCL) là vùng đất thấp ven biển của Việt Nam và là khu vực bị tác hại nặng nề nhất do biến đổi khí hậu gây ra. Thành phố Cần Thơ nằm ở trung tâm ĐBSCL với đặc điểm là nắng nhiều và nhiệt độ cao quanh năm. Mùa mưa kéo dài từ tháng 5 đến tháng 10, mùa khô từ tháng 11 đến tháng 4 năm sau. Ngoài ra do nằm cạnh sông Hậu nên Cần Thơ có mạng lưới sông, kênh, rạch khá chằng chịt. Vùng tứ giác Long Xuyên có địa hình thấp trũng và chịu ảnh hưởng lũ trực tiếp hàng năm. Theo báo cáo của Bộ Tài nguyên và Môi trường năm 2011 thì trị số phổ biến của lượng bức xạ tổng cộng trung bình năm là 150-170 kcal/cm2 và trị số phổ biến về lượng mưa trung bình năm khoảng 1600 đến 2000 mm (Bộ Tài nguyên và Môi trường, 2011). Lượng mưa ngày lớn nhất ở Thành phố Cần Thơ khoảng 150-350 mm. Cả mùa mưa có từ 4 đến 6 tháng mưa trên 200 mm/tháng. Việc biến đổi khí hậu sẽ làm thiệt hại cho sản xuất nông nghiệp do đất đai bị bạc màu và nhiễm mặn, hạn hán bất thường, lũ lụt không theo qui luật và nhiều dịch bệnh mới hình thành,...

Nghiên cứu về tác động của biến đổi khí hậu đối với tài nguyên nước đòi hỏi phải chi tiết hóa (downscaling) lượng mưa hằng ngày từ các dự báo cấp khu vực (Regional Climate Model - RCM). Bài báo này đề xuất một phương pháp downscaling hai bước để dự báo lượng mưa hằng ngày. Bước đầu tiên thực hiện việc dự báo một ngày nào đó có mưa hay không. Bước thứ hai sẽ dự báo lượng mưa nếu như ngày đó được dự báo là có mưa ở bước một.

Các phần tiếp theo của bài báo này như sau: phần 2 trình bày ngắn gọn về các nghiên cứu liên quan đến mô hình dự báo lượng mưa, phần 3 trình bày các mô hình dự báo lượng mưa. Phần 4 trình bày các kết quả thực nghiệm, tiếp theo sau đó là phần kết luận và hướng phát triển.

2 NGHIÊN CỨU LIÊN QUAN

Các chuyên gia đã sử dụng mô hình tuần hoàn tổng quát (GCM - General Circulation Model) để thiết kế mô hình và mô phỏng các tiến trình biến đổi khí hậu trong phạm vi toàn cầu (Ghosh *et al*., 2008). Mô hình GCM sử dụng các biến thời tiết có độ phân giải thấp để dự báo các biến đổi khí hậu dài hạn và trung hạn cho các vùng với phạm vi rộng lớn, do đó làm cho các chuyên gia khó khăn trong việc dự báo ảnh hưởng của biến đổi khí hậu đối với nguồn tài nguyên nước tại các vùng có phạm vi nhỏ. Việc biến đổi kết quả đầu ra của mô hình GCM để dự báo biến đổi khí hậu tại các vùng có phạm vi nhỏ hơn (như: cấp xã, ấp, cánh đồng) là một bài toán khó vì mô hình GCM không đề cập đến các tiến trình cơ bản xảy ra ở các vùng có phạm vi nhỏ (ví dụ: tiến trình bốc hơi nước, hấp thụ nước, phân bố lượng mưa).

Các phương pháp downscaling đã được phát triển để tạo sự liên hệ giữa kết quả đầu ra của mô hình GCM có độ phân giải thấp với các biến thời tiết có độ phân giải cao hơn ở các vùng có phạm vi nhỏ. Các phương pháp downscaling có thể được phân thành hai nhóm chính: downscaling thống kê và downscaling động. Phương pháp downscaling thống kê có thể được chia thành bốn nhóm: phân loại thời tiết (weather typing method) (Bárdossy *et al*., 1992; Von Storch *et al*., 1993; Bárdossy , 1997), bộ sinh dữ liệu thời tiết ngẫu nhiên (stochastic weather generator) (Selker and Haith, 1990; Tung and Haith, 1995; Yu *et al*., 2002), phương pháp lấy mẫu lại (resampling method) (Murphy, 2000; Buishand and Brandsma, 2001; Palutikor *et al*., 2002) và phương pháp hồi quy (regression method).

Phương pháp hồi quy thiết lập một hàm tuyến tính hoặc phi tuyến thực nghiệm giữa các biến thời tiết ở cấp độ vùng có phạm vi nhỏ (cấp độ địa phương-local scale) và các biến ở cấp độ toàn cục (global scale) của mô hình GCM. Phương pháp này thường được sử dụng vì dễ cài đặt. Ngoài ra, hàm hồi quy cho downscaling có thể được xây dựng bằng mạng nơ-ron (Neural network) (Hewitson and Crane, 1996; Olsson *et al*., 2001; Dibike and Coulibaly, 2006), phân tích tương quan chính tắc (Burger, 1996; Menzel and Burger, 2002; Chu *et al*., 2008) hay máy học véc-tơ hỗ trợ (Support vector machine) (Tripathi *et al*., 2006; Anamdhi *et al*., 2008). Nghiên cứu của (Chen *et al*., 2010) đề xuất kết hợp mô hình phân lớp (mưa hay không mưa) và mô hình hồi quy sử dụng máy học véc tơ hỗ trợ.

Nhiều mô hình và phần mềm downscaling đã được hình thành và phát triển. Nhưng mô hình SDSM (Statistical downscaling model) của Wilby *et al*. (2002) được sử dụng nhiều nhất. Ví dụ như, Wilby *et al*. (2006) đã kết hợp SDSM với một mô hình cân bằng nước và mô hình chất lượng nước cân bằng để nghiên cứu đánh giá ảnh hưởng của

biến đổi khí hậu và sự không chắc chắn trong các dòng chảy của sông. Thêm vào đó, SDSM thường được so sánh với các phương pháp downscaling thống kê.

Trong nghiên cứu dự báo lượng mưa, chúng tôi trước tiên sử dụng mô hình hồi quy tuyến tính. Tiếp đến, nghiên cứu tập trung vào hướng tiếp cận dựa trên các mô hình máy học tự động như: k láng giềng (k Nearest Neighbors) (Fix and Hodges, 1952), cây quyết định (Decision Trees) (Breiman *et al*., 1984), bagging (Breiman, 1996), rừng ngẫu nhiên (Random Forests) (Breiman, 2001) và máy học véc tơ hỗ trợ (Support Vector Machines) (Vapnik, 1995). Chúng tôi cũng đề xuất mô hình học phân cấp kết hợp giữa mô hình phân lớp và mô hình hồi quy dựa trên rừng ngẫu nhiên và máy học véc tơ hỗ trợ.

3 MÔ HÌNH DỰ BÁO

3.1 Mô hình hồi quy tuyến tính (linear regression - LM)

Hồi quy là phương pháp toán học được áp dụng thường xuyên trong thống kê để phân tích mối liên hệ giữa các hiện tượng kinh tế xã hội. Hồi quy tuyến tính được sử dụng rộng rãi trong thực tế do tính chất đơn giản hóa của hồi quy.

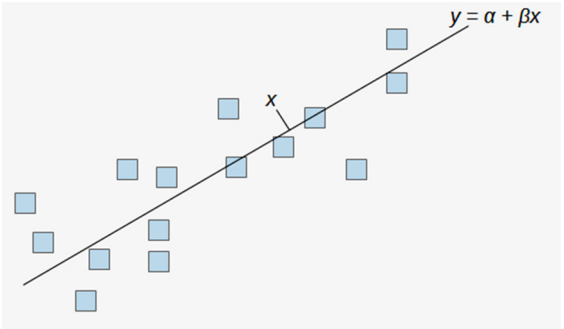
Phân tích hồi quy là phân tích thống kê để xác định mối quan hệ giữa biến phụ thuộc y với một hay nhiều biến độc lập x. Mô hình hồi quy đơn giản nhất là hàm tuyến tính (bậc 1) dùng để mô tả mối quan hệ giữa biến phụ thuộc và biến độc lập là tuyến tính. Mô hình hồi quy tuyến tính có dạng:

$$y = \alpha + \beta x \quad (1)$$

với α là chặn (intercept), β là độ dốc (slope)

Các tham số α, β của mô hình được ước lượng từ dữ liệu quan sát. Xét tập dữ liệu gồm m phần tử $x_1, x_2, \ldots, x_m$ trong không gian n chiều (biến độc lập, thuộc tính), có giá trị tương ứng của biến phụ thuộc (cần dự báo) là $y_1, y_2, \ldots, y_m$. Các tham số α, β của mô hình được ước lượng bằng phương pháp bình phương bé nhất (least squares):

$$Min\left(\sum_{i=1}^{m}[y_i - (\alpha + \beta x_i)]^2\right) \quad (2)$$



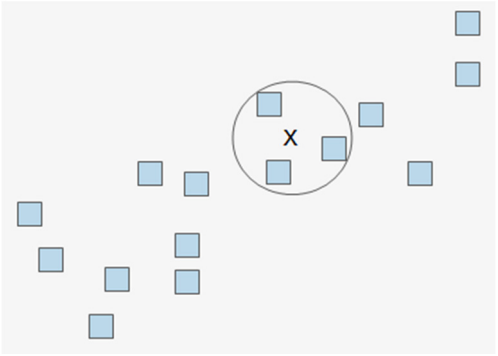
Hình 1: Hồi quy tuyến tính

Giá trị dự báo cho phần tử mới x dựa vào công thức (3):

$$\hat{y} = \alpha + \beta x \quad (3)$$

3.2 k láng giềng (k Nearest Neighbors - kNN)

Giải thuật k láng giềng (kNN) được Fix và Hodges đề xuất từ những năm 1952. Đây là phương pháp rất đơn giản nhưng cũng cho hiệu quả cao trong khai mỏ dữ liệu (Hastie *et al*., 2009; Wu and Kumar, 2009). Giả sử có tập dữ liệu bao gồm m phần tử $x_1, x_2, \ldots, x_m$ trong không gian n chiều, có giá trị tương ứng của biến phụ thuộc là $y_1, y_2, \ldots, y_m$.



Hình 2: Giải thuật k láng giềng

Giải thuật kNN không có quá trình học. Khi dự đoán giá trị biến phụ thuộc của phần tử dữ liệu x mới đến, giải thuật đi tìm k láng giềng ($k=1, 2, \ldots$) của x từ tập dữ liệu học là các phần tử $\{(x_1,y_1), \ldots, (x_k,y_k)\}$, sau đó thực hiện:

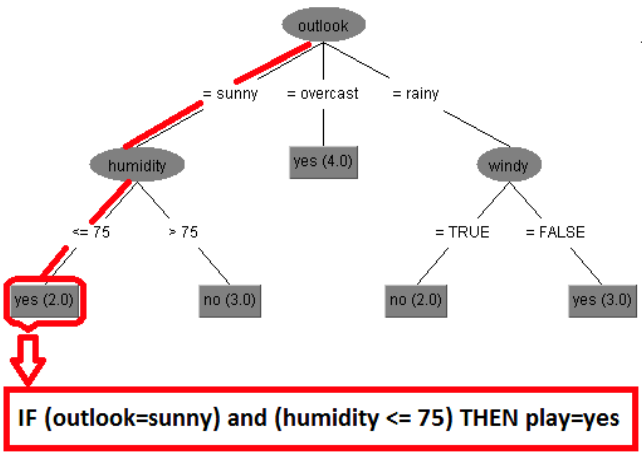
– Phân lớp với bình chọn số đông trong các giá trị $\{y_1, \ldots, y_k\}$,

– Hồi quy với giá trị trung bình của các $\{y_1, \ldots, y_k\}$.

Quá trình tìm k láng giềng của $\boldsymbol{x}$ thường sử dụng khoảng cách (distance) hay độ tương tự (similarity).

3.3 Cây quyết định (Decision Trees - DT)

Cây quyết định đề xuất bởi (Breiman *et al.*, 1984; Quinlan, 1993) là mô hình máy học tự động sử dụng rất nhiều trong khai mỏ dữ liệu (Wu and Kumar, 2009) do tính đơn giản và hiệu quả. Hình 3 minh họa một ví dụ của cây quyết định thu được bằng cách học từ tập dữ liệu, để dự đoán chơi Golf (ŷ = yes / no?) từ các biến (thời tiết, nhiệt độ, độ ẩm, gió). Mô hình rất dễ hiểu bởi vì chúng ta có thể rút trích luật quyết định tương ứng với nút lá có dạng IF-THEN được tạo ra từ việc thực hiện AND trên các điều kiện theo đường dẫn từ nút gốc đến nút lá. Các luật quyết định dễ hiểu với người sử dụng.



outlook	temp.	hum.	windy	Play
sunny	85	85	false	No
sunny	80	90	true	No
overcast	83	78	false	Yes
rain	70	96	false	Yes
rain	68	80	false	Yes
rain	65	70	true	No
. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .

Hình 3: Cây quyết định học từ dữ liệu cho phép dự báo chơi Golf

Xét tập dữ liệu bao gồm $\boldsymbol{m}$ phần tử $\boldsymbol{x_1, x_2, \ldots, x_m}$ trong không gian $\boldsymbol{n}$ chiều, có giá trị tương ứng của biến phụ thuộc là $\boldsymbol{y_1, y_2, \ldots, y_m}$. Giải thuật học từ dữ liệu là quá trình xây dựng cây bắt đầu từ nút gốc đến nút lá. Đây là giải thuật đệ quy phân hoạch tập dữ liệu theo các biến độc lập thành các phân vùng chữ nhật rời nhau mà ở đó các phần tử dữ liệu $\boldsymbol{x_i}$, $\boldsymbol{x_j}$, …, $\boldsymbol{x_k}$ của cùng phân vùng (nút lá) có các $\boldsymbol{y_i}$, $\boldsymbol{y_j}$, …, $\boldsymbol{y_k}$ là thuần khiết:

- Giống nhau trong vấn đề phân lớp,
- Tương tự nhau trong vấn đề hồi quy.

Giải thuật học mô hình cây quyết định từ dữ liệu gồm 2 bước lớn: xây dựng cây, cắt nhánh để tránh học vẹt. Quá trình xây dựng cây được làm như sau:

- Bắt đầu từ nút gốc, tất cả các dữ liệu học ở nút gốc,
- Nếu các phần tử dữ liệu tại 1 nút là thuần khiết thì nút đang xét được cho là nút lá, giá trị dự báo của nút lá cho vấn đề phân lớp với bình chọn số đông trong các giá trị $\{\boldsymbol{y_i, \ldots, y_k}\}$, cho vấn đề hồi quy với giá trị trung bình của các $\{\boldsymbol{y_i, \ldots, y_k}\}$.
- Nếu dữ liệu ở nút quá hỗn loạn (các giá trị $\{\boldsymbol{y_i, \ldots, y_k}\}$ rất khác nhau) thì nút được cho là nút trong, tiến hành phân hoạch dữ liệu một cách đệ quy bằng việc chọn 1 biến để thực hiện phân hoạch tốt nhất có thể.

Một biến được cho là tốt được sử dụng để phân hoạch dữ liệu sao cho kết quả thu được cây nhỏ nhất. Việc lựa chọn này dựa vào các heuristics: chọn biến sinh ra các nút thuần khiết nhất. Hiện nay có 2 giải thuật học cây quyết định tiêu biểu là C4.5 của (Quinlan, 1993), CART của (Breiman *et al.*, 1984).

Để đánh giá và chọn biến khi phân hoạch dữ liệu, Quinlan đề nghị sử dụng độ lợi thông tin (chọn biến có độ lợi thông tin lớn nhất) và tỉ số độ lợi dựa trên hàm entropy của Shannon. Độ lợi thông tin của một biến được tính bằng: độ đo hỗn loạn trước khi phân hoạch trừ cho sau khi phân hoạch. Giả sử $\boldsymbol{p_c}$ là xác suất mà phần tử trong dữ liệu $\boldsymbol{D}$ thuộc lớp $\boldsymbol{y_c}$ (c =1 , C), độ đo hỗn loạn thông tin trước khi phân hoạch được tính theo công thức entropy (4) như sau:

$$Info(D) = -\sum_{c=1}^{C} p_c \log_2 p_c \qquad (4)$$

Độ đo hỗn loạn sau khi sử dụng biến ***A*** phân hoạch dữ liệu ***D*** có ***m*** phần tử thành ***v*** phân vùng kích thước tương ứng là m_1, m_2, ..., m_v được tính bởi (5):

$$Info_A(D) = -\sum_{j=1}^{v} \frac{m_j}{m} Info(D_j) \quad (5)$$

Độ lợi thông tin khi chọn biến ***A*** phân hoạch dữ liệu ***D*** thành ***v*** phần được tính bởi công thức (6):

$$Gain(A) = Info(D) - Info_A(D) \quad (6)$$

Giải thuật CART của Breiman và các cộng sự sử dụng chỉ số Gini để phân hoạch dữ liệu trong quá trình xây dựng cây. Giả sử p_c là xác suất mà phần tử trong dữ liệu ***D*** thuộc lớp y_c (c =1, C), chỉ số Gini được tính theo công thức (7):

$$Gini\ (D) = 1 - \sum_{c=1}^{k} p_c^2 \quad (7)$$

Hàm Gini nhỏ nhất khi lớp trong ***D*** bị lệch. Nếu sử dụng biến ***A*** phân hoạch ***D*** kích thước ***m*** thành 2 tập con D_1 (kích thước m_1) và D_2 (kích thước m_2), hàm Gini được tính như công thức (8). Biến được chọn phân hoạch dữ liệu là biến cho giá trị chỉ số Gini nhỏ nhất.

$$Gini_A(D) = \frac{m_1}{m} Gini(D_1) + \frac{m_2}{m} Gini(D_2) \quad (8)$$

Cho vấn đề hồi quy, độ đo hỗn loạn thông tin tại phân vùng ***D*** dựa trên độ lệch chuẩn như trong (9) với μ là giá trị trung bình của các giá trị ***y*** trong ***D***.

$$S(D) = \sum_{i=1}^{k} \frac{(y_i - \mu)^2}{k} \quad (9)$$

Nếu sử dụng biến ***A*** phân hoạch ***D*** kích thước ***m*** thành 2 tập con D_1 (kích thước m_1) và D_2 (kích thước m_2), độ hỗn loạn sau khi phân hoạch được tính như công thức (10).

$$S_A(D) = \frac{m_1}{m} S(D_1) + \frac{m_2}{m} S(D_2) \quad (10)$$

Biến được chọn phân hoạch dữ liệu là biến cho giá trị độ hỗn loạn trước khi phân hoạch trừ cho độ hỗn loạn sau khi phân hoạch là nhỏ nhất.

Mô hình cây quyết định sau khi xây dựng thường không mạnh với nhiễu và dễ dẫn đến học vẹt. Tức là mô hình có tính tổng quát thấp, chỉ cần dữ liệu kiểm tra có thay đổi một ít so với dữ liệu học thì cây quyết định dự báo sai. Để khắc phục khuyết điểm này, Quinlan cũng đề nghị các chiến lược cắt nhánh trong giải thuật C4.5. Có 2 lựa chọn hoặc postpruning (cắt nhánh cây sau khi xây dựng cây) hay prepruning (dừng sớm quá trình phân nhánh). Trong thực tế, postpruning được sử dụng nhiều hơn prepruning. Tuy nhiên độ phức tạp của việc cắt nhánh sau khi xây dựng cây rất phức tạp, sử dụng các chiến lược để ước lượng lỗi sinh ra bởi mô hình sau khi cắt nhánh.

3.4 Mô hình Bagging (BagDT)

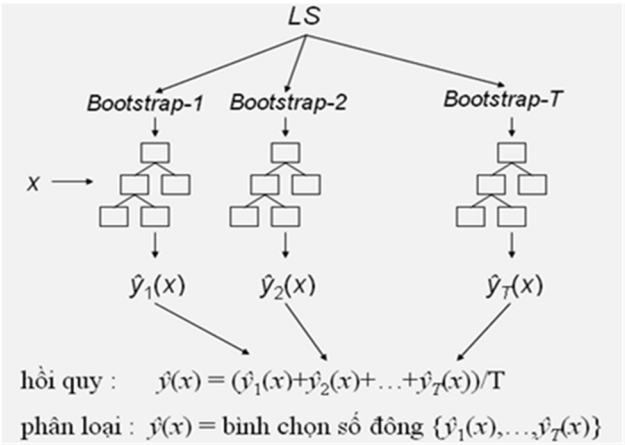
Từ những năm 1990, cộng đồng máy học đã nghiên cứu cách để kết hợp nhiều mô hình phân loại yếu thành mô hình tập hợp phân loại mạnh cải thiện độ chính xác cao hơn so với chỉ một mô hình phân loại đơn yếu. Trong phân tích thành phần lỗi của giải thuật học, Breiman đã chỉ ra trong (Breiman, 1996), lỗi bao gồm 2 thành phần là bias và variance. Thành phần lỗi bias là khái niệm về lỗi của mô hình học (không liên quan đến dữ liệu học) và thành phần lỗi variance là lỗi do tính biến thiên của mô hình so với tính ngẫu nhiên của các mẫu dữ liệu học. Mục đích của các mô hình tập hợp là làm giảm variance và/hoặc bias của các giải thuật học. Dựa trên cách phân tích hiệu quả của giải thuật học dựa trên thành phần lỗi bias và variance, Breiman đã đề xuất giải thuật học Bagging (Bootstrap AGGregatING) nhằm giảm lỗi variance của giải thuật học nhưng không làm tăng lỗi bias quá nhiều. Giải thuật có thể được tóm tắt như sau:

– Từ tập dữ liệu học ***LS*** có ***m*** phần tử, xây dựng ***T*** mô hình cơ sở độc lập nhau.

– Mô hình thứ ***t*** được xây dựng trên tập mẫu Bootstrap thứ ***t*** (lấy mẫu ***m*** phần tử có hoàn lại từ tập học ***LS***).

– Kết thúc quá trình xây dựng ***T*** mô hình cơ sở, dùng chiến lược bình chọn số đông để phân lớp một phần tử ***x*** mới đến hoặc giá trị trung bình cho bài toán hồi quy.

Trong thực tế, giải thuật Bagging cải thiện rất tốt các mô hình đơn không ổn định như cây quyết định và thường có thành phần lỗi variance cao. Hình 4 là ví dụ của giải thuật Bagging được áp dụng cho mô hình cơ sở là cây quyết định.



Hình 4: Giải thuật Bagging của cây quyết định

3.5 Rừng ngẫu nhiên (Random Forests - RF)

Tiếp cận rừng ngẫu nhiên do (Breiman, 2001) đưa ra là một trong những phương pháp tập hợp mô hình thành công nhất. Giải thuật rừng ngẫu nhiên tạo ra một tập hợp các cây quyết định không cắt nhánh, mỗi cây được xây dựng trên tập mẫu bootstrap (như Bagging), tại mỗi nút phân hoạch tốt nhất được thực hiện từ việc chọn ngẫu nhiên một tập con các thuộc tính. Lỗi tổng quát của rừng phụ thuộc vào độ chính xác của từng cây thành viên trong rừng và sự phụ thuộc lẫn nhau giữa các cây thành viên. Giải thuật rừng ngẫu nhiên xây dựng cây không cắt nhánh nhằm giữ cho thành phần lỗi bias thấp và dùng tính ngẫu nhiên để điều khiển tính tương quan thấp giữa các cây trong rừng. Tiếp cận rừng ngẫu nhiên cho độ chính xác cao khi so sánh với các thuật toán học có giám sát hiện nay. Như Breiman đã đề cập trong (Breiman, 2001), rừng ngẫu nhiên học nhanh, chịu đựng nhiễu tốt và không bị tình trạng học vẹt. Giải thuật rừng ngẫu nhiên sinh ra mô hình có độ chính xác cao đáp ứng được yêu cầu thực tiễn cho vấn đề phân loại, hồi qui. Giải thuật rừng ngẫu nhiên (Hình 5) có thể được trình bày ngắn gọn như sau:

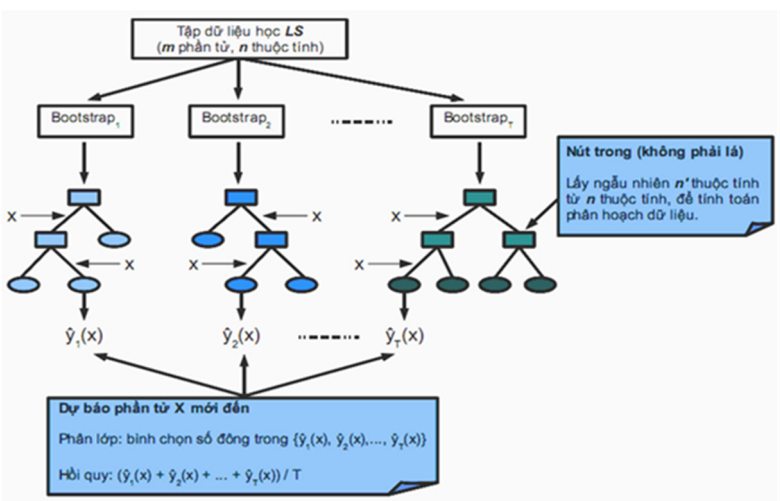
– Từ tập dữ liệu học LS có m phần tử và n biến (thuộc tính), xây dựng T cây quyết định một cách độc lập nhau.

– Mô hình cây quyết định thứ t được xây dựng trên tập mẫu Bootstrap thứ t (lấy mẫu m phần tử có hoàn lại từ tập học LS).

– Tại nút trong, chọn ngẫu nhiên n' biến (n'<<n) và tính toán phân hoạch tốt nhất dựa trên n' biến này.

– Cây được xây dựng đến độ sâu tối đa không cắt nhánh.

Kết thúc quá trình xây dựng T mô hình cơ sở, dùng chiến lược bình chọn số đông để phân lớp một phần tử mới đến hoặc giá trị trung bình cho bài toán hồi quy.

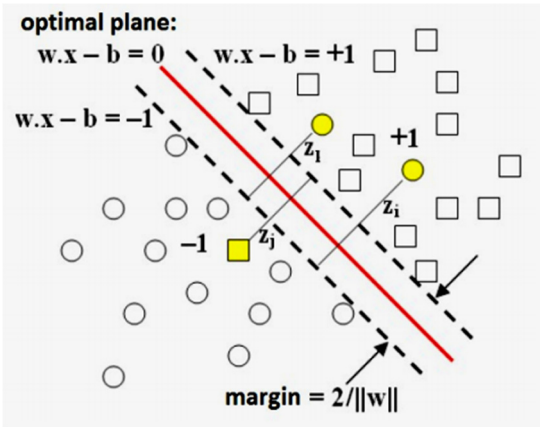


Hình 5: Giải thuật rừng ngẫu nhiên

3.6 Máy học véc tơ hỗ trợ (Support Vector Machines - SVM)

Máy học véc tơ hỗ trợ (SVM) được đề xuất bởi (Vapnik, 1995) là mô hình hiệu quả và phổ biến cho vấn đề phân lớp, hồi quy những tập dữ liệu có số chiều lớn.

Xét ví dụ phân lớp nhị phân tuyến tính (hình 6) với m phần tử $x_1, x_2, \ldots, x_m$ trong không gian n chiều, có nhãn (lớp) của các phần tử là $y_1, y_2, \ldots, y_m$ có giá trị 1 hoặc -1. $y_i = 1$, nếu x_i thuộc lớp $+1$ (lớp dương, lớp chúng ta quan tâm), $y_i = -1$, nếu x_i thuộc lớp -1 (lớp âm hay các lớp còn lại). SVM tìm siêu phẳng tối ưu (xác định bởi véc tơ pháp tuyến w và độ lệch của siêu phẳng b) dựa trên 2 siêu phẳng hỗ trợ của 2 lớp. Các phần tử lớp +1 nằm bên phải của siêu phẳng hỗ trợ cho lớp +1, các phần tử lớp -1 nằm phía bên trái của siêu phẳng hỗ trợ cho lớp -1. Những phần tử nằm ngược phía với siêu phẳng hỗ trợ được coi như lỗi. Khoảng cách lỗi được biểu diễn bởi $z_i \geq 0$ (với x_i nằm đúng phía của siêu phẳng hỗ trợ của nó thì khoảng cách lỗi tương ứng $z_i = 0$, còn ngược lại thì $z_i > 0$ là khoảng cách từ điểm x_i đến siêu phẳng hỗ trợ tương ứng của nó).



Hình 6: Phân lớp tuyến tính với máy học véc tơ hỗ trợ

Khoảng cách giữa 2 siêu phẳng hỗ trợ được gọi là lề. Siêu phẳng tối ưu (nằm giữa 2 siêu phẳng hỗ trợ) tìm được từ 2 tiêu chí là cực đại hóa lề (lề càng lớn, mô hình phân lớp càng an toàn) và cực tiểu hóa lỗi. Vấn đề dẫn đến việc giải bài toán quy hoạch toàn phương (11):

$$\min \Psi(w, b, z) = (1/2)\,\|w\|^2 + c\sum_{i=1}^{m} z_i \quad s.t \qquad (11)$$

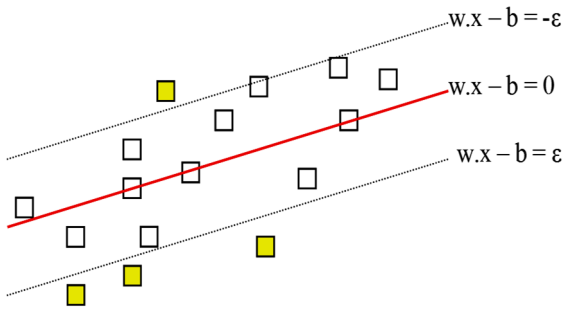
$$y_i(w.x_i - b) + z_i \geq 1$$

$$z_i \geq 0 \quad (i=1, 2, \ldots, m)$$

với hằng c > 0 được sử dụng để chỉnh độ rộng lề và lỗi.

Giải bài toán quy hoạch toàn phương (11), thu được (w, b). Phân lớp phần tử x dựa vào dấu của $(w.x - b)$ như trong (12). Nếu giá trị biểu thức $(w.x - b) > 0$ thì gán nhãn cho x là lớp dương (+1), ngược lại thì gán nhãn cho x là lớp âm (-1).

$$\text{predict}(x) = \text{sign}(w.x - b) \qquad (12)$$



Hình 7: Hồi quy với máy học véc tơ hỗ trợ

Máy học SVM cũng có thể xử lý bài toán hồi quy. Trong vấn đề hồi quy như Hình 7, SVM tìm siêu phẳng đi qua tất cả các phần tử dữ liệu với độ lệch chuẩn là ε. Huấn luyện máy học SVM cho xử lý vấn đề hồi quy dẫn đến việc giải bài toán quy hoạch toàn phương (13) như sau:

$$\min \Psi(w, b, z^*, z) = (1/2)\,\|w\|^2 + c\sum_{i=1}^{m}(z_i^* + z_i) \quad s.t. \qquad (13)$$

$$w.x_i - b - y_i - z_i^* \leq \varepsilon$$

$$w.x_i - b - y_i + z_i \geq -\varepsilon$$

$$z_i^*, z_i \geq 0 (i=1, 2, \ldots, m)$$

với hằng c > 0 được sử dụng để chỉnh độ rộng lề và lỗi.

Giải bài toán quy hoạch toàn phương (13) sẽ thu được siêu phẳng hồi quy (w, b) của SVM. Dự báo cho phần tử mới đến x dựa trên siêu phẳng (w, b) được tính theo công thức (14):

$$\text{predict}(x) = (w.x - b) \qquad (14)$$

Giải thuật SVM có thể thay thế các tích vô hướng trong các công thức (11-14) bởi hàm nhân (kernel functions), sẽ cho phép giải quyết một số lớn các bài toán phân lớp và hồi quy phi tuyến. Không có bất kỳ một thay đổi nào cần thiết về mặt

giải thuật, việc làm duy nhất là thay thế các tích vô hướng của hai véc tơ trong các công thức bởi một trong các hàm nhân cơ bản được dùng phổ biến như:

- Đa thức bậc d: $K(u, v) = (u.v + c)^d$ (15)
- Radial Basis Function (RBF):

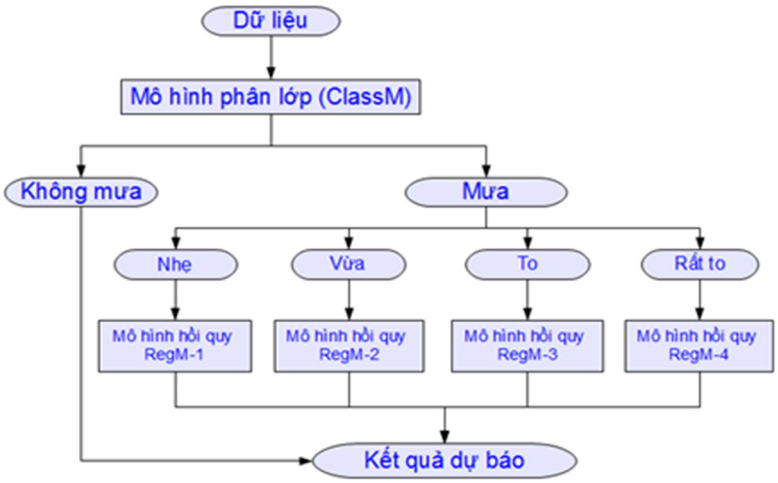
$K(u, v) = exp(-\gamma||u - v||^2)$ (16)

3.7 Mô hình hồi quy phân cấp

Chúng ta có thể sử dụng trực tiếp các mô hình hồi quy vừa được trình bày để dự báo lượng mưa trong ngày. Mỗi mô hình đều có ưu điểm và khuyết điểm khác nhau. Chẳng hạn mô hình hồi quy tuyến tính thì rất đơn giản, thời gian xây dựng mô hình và dự báo nhanh, điều tất yếu là độ chính xác cũng không cao. Riêng mô hình kNN cũng đơn giản, chỉ sử dụng duy nhất tham số là ***k = 1, 2, ...*** là số láng giềng, tuy nhiên thời gian dự báo lâu hơn do phải tìm kiếm láng giềng của phần tử cần dự báo. Mô hình cây quyết định chỉ cần duy nhất tham số ***minobj = 1,2, ...*** là số phần tử tối thiểu tại mỗi nút lá, thời gian xây dựng mô hình và dự báo nhanh, đạt được độ chính xác tương đối cao so với kNN và hồi quy tuyến tính. Bagging và rừng ngẫu nhiên thì cần thêm tham số là số lượng cây ***T = 50, 100,...*** riêng, rừng ngẫu nhiên còn sử dụng thêm tham số là số biến ngẫu nhiên sử dụng cho phân hoạch (trong khoảng $\left[\sqrt{n}, \frac{n}{2}\right]$ với ***n*** là số biến của dữ liệu). Cho dù phức tạp, nhưng Bagging, rừng ngẫu nhiên vẫn đơn giản hơn khi so với máy học SVM. Xây dựng mô hình SVM cần thiết ba tham số là hằng số ***c > 0*** (để chỉnh độ rộng lề và lỗi), độ lệch chuẩn là ***ε*** và tham số của hàm nhân. Thời gian xây dựng mô hình và dự báo rất cao (ít nhất là bậc 2 so với số lượng phần tử). Mặc dù phức tạp, nhưng Bagging, rừng ngẫu nhiên và SVM là mô hình phi tuyến, nên xử lý tốt cho các vấn đề phi tuyến, đặc biệt là dự báo lượng mưa đang xét ở đây.

Hình 8 minh họa mô hình hồi quy phân cấp. Dữ liệu được phân lớp (ClassM) vào một trong năm lớp như: không mưa (lượng mưa = 0), mưa nhẹ (lượng mưa: 0-2,5 mm), mưa vừa (lượng mưa: 2,5-7,6 mm), mưa to (lượng mưa: 7,6-50mm), rất to (lượng mưa trên 50 mm). Tương ứng với từng lớp, một mô hình hồi quy được xây dựng cho phép dự báo tốt các phần tử thuộc lớp đó (RegM-i).

Xét về độ phức tạp, xử lý vấn đề phân lớp đơn giản hơn rất nhiều so với bài toán hồi quy. Hơn nữa, quá trình xây dựng mô hình hồi quy càng phức tạp hơn khi cần dự báo lượng mưa từ tập dữ liệu, có mối quan hệ phi tuyến giữa biến phụ thuộc (lượng mưa) với nhiều biến độc lập (bức xạ mặt trời, hướng gió, tốc độ gió, nhiệt độ). Từ phân tích trên, chúng tôi đề xuất mô hình hồi quy phân cấp, kết hợp giữa mô hình phân lớp và nhiều mô hình hồi quy cục bộ để nâng cao hiệu quả xử lý của dự báo lượng mưa.



Hình 8: Mô hình phân cấp (phân lớp + hồi quy)

4 KẾT QUẢ THỰC NGHIỆM

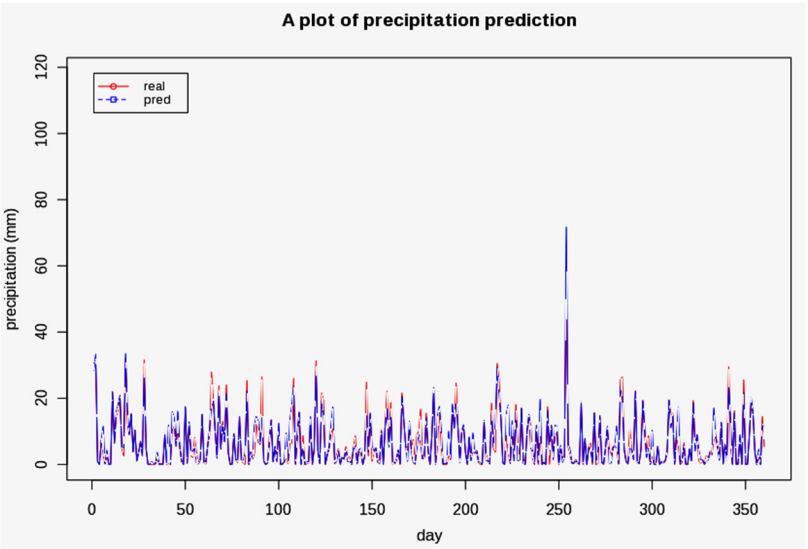
Để tiến hành đánh giá hiệu quả của các mô hình dự báo lượng mưa, chúng tôi tiến hành cài đặt tất cả các chương trình dự báo bằng ngôn ngữ R (Ihaka and Gentleman, 1996) có sử dụng các gói thư viện FNN, rpart, ipred, randomForest, e1071. Chương trình bao gồm các mô hình: Hồi quy tuyến tính (LM), k láng giềng (kNN), Cây quyết định (DT), Bagging (BagDT), Rừng ngẫu nhiên (RF), Máy học véc tơ hỗ trợ cho hồi quy SVR, Mô hình phân cấp: RF phân lớp và RF hồi quy (RFC-RFR), Mô hình phân cấp: SVC phân lớp và SVR hồi quy (SVC-SVR) để dự báo lượng mưa.

Bảng 1: Kết quả dựa báo lượng mưa của các mô hình

Phương pháp	MSE	MAE
hồi quy tuyến tính (lm)	34.161274	4.406217
k láng giềng (k=5)	19.815441	2.593839
cây quyết định (leaf-size=5)	15.508522	2.052494
Bagging (#trees=100)	**8.970759**	1.481438
rừng ngẫu nhiên (#trees=100, #randim=3)	9.131517	***1.545566***
máy học SVR (RBF, γ=0.01, ε=0.1, C= 10^4)	17.006430	2.455377
mô hình phân cấp RFC-RFR	***10.412143***	**1.469576**
mô hình phân cấp SVC-SVR	20.074542	2.405764

Chúng tôi sử dụng tập dữ liệu (gồm 24 tập con) từ SEA-START RC có địa chỉ website là http://cc.start.or.th. Đây là hệ thống phân phối dữ liệu biến đổi khí hậu nằm trong chương trình hợp tác giữa trung tâm START khu vực Đông Nam Á và ESRI của Thái Lan. Tập dữ liệu thu được từ kết quả của quá trình mô phỏng phức tạp cho từng ngày với kích thước lưới là 20 x 20 km cho toàn bộ khu vực sông Mêkong trong khoảng từ năm 1980 đến năm 2006. Chúng tôi chỉ sử dụng dữ liệu ở lưới gần Cần Thơ (có kinh độ LON = 105.8 và vĩ độ LAT = 10.2). Tập dữ liệu có 9360 dòng (ngày), mỗi dòng có 6 giá trị thuộc tính là nhiệt độ (tmax, tmin), bức xạ mặt trời (solar radiation), hướng gió (wind-dir), tốc độ gió (wind-speed) và lượng mưa (rainfall). Vấn đề chúng ta cần kiểm thử là xây dựng các mô hình dự báo sử dụng tập dữ liệu có được để dự báo lượng mưa (rainfall) từ 5 thuộc tính còn lại. Chúng tôi sử dụng nghi thức kiểm thử hold-out bằng cách lấy ngẫu nhiên 2/3 tập dữ liệu (6240 dòng) làm tập huấn luyện các mô hình dự báo và 1/3 còn lại (3120 dòng) làm tập kiểm tra kết quả dự báo. Kết quả dự báo được đánh giá trên tiêu chí trung bình bình phương lỗi (Mean Square Error - MSE) và trung bình lỗi tuyệt đối (Mean Absolute Error - MAE). Chúng tôi chỉ sử dụng tập huấn luyện để điều chỉnh các tham số của các mô hình. Các tham số này được lựa chọn sao cho đạt tiêu chí lỗi thấp nhất.

Kết quả thu được từ các mô hình dự báo (với các tham số tối ưu) được trình bày trong Bảng 1. Ở hai cột **MSE** và **MAE**, kết quả dự báo với lỗi thấp nhất được in đậm, lỗi thấp thứ hai được in gạch dưới và lỗi thấp thức ba được in đậm và nghiêng.



Hình 9: Kết quả dự báo 360 ngày của mô hình phân cấp RFC-RFR

Không có gì ngạc nhiên khi mô hình hồi quy tuyến tính cho lỗi dự báo cao nhất. Trong khi các mô hình dự báo phi tuyến chứng tỏ nhiều ưu thế hơn. Mặc dù vậy, mô hình máy học véc tơ hỗ trợ cho hồi quy SVR và cả mô hình phân cấp SVC-SVR vẫn chỉ thắng thế khi so sánh với kNN và hồi quy tuyến tính. Trong khi đó, mô hình cây quyết định đơn giản cũng cho kết quả rất khả quan khi so sánh với tất cả các mô hình còn lại. Tuy nhiên, hiệu quả nhất vẫn là phương pháp tập hợp mô hình như Bagging, rừng ngẫu nhiên và mô hình phân cấp RFC-RFR, cho phép dự báo rất chính xác lượng mưa (lỗi dự báo thấp). Mô hình Bagging dự báo

với trung bình bình phương lỗi nhỏ nhất trong khi mô hình phân cấp RFC-RFR có thể dự báo với trung bình lỗi tuyệt đối là nhỏ nhất.

Đồ thị về kết quả dự báo lượng mưa của 360 ngày của mô hình phân cấp RFC-RFR được trình bày trong Hình 9. Quan sát đồ thị này, chúng ta có thể thấy rằng mô hình phân cấp RFC-RFR dự báo hiệu quả lượng mưa.

5 KẾT LUẬN VÀ ĐỀ XUẤT

Nghiên cứu này đã so sánh các mô hình dự báo theo phương pháp phân tích lỗi dự báo. Phương pháp downscaling hai bước được đề xuất trong bài báo này nhằm dự báo lượng mưa hằng ngày và cho thấy khả năng ứng dụng của các mô hình dự báo lượng mưa trong thực tế. Nghiên cứu ngày đã áp dụng các phương pháp Hồi quy tuyến tính, k láng giềng, Cây quyết định, Bagging, Rừng ngẫu nhiên (RF), Máy học véc tơ hỗ trợ cho hồi quy SVR, Mô hình phân cấp: RF phân lớp và RF hồi quy (RFC-RFR), Mô hình phân cấp: SVC phân lớp và SVR hồi quy (SVC-SVR) để dự báo lượng mưa từ tập dữ liệu của SEA-START ở lưới gần Thành phố Cần Thơ. Kết quả thực nghiệm cho thấy rằng mô hình hồi quy tuyến tính không phù hợp cho dự báo lượng mưa trong khi các mô hình dự báo khác như Bagging, rừng ngẫu nhiên và mô hình phân cấp RFC-RFR dự báo chính xác hơn.

Trong tương lai, chúng tôi sẽ áp dụng các mô hình dự báo này vào dữ liệu thực tế của Thành phố Cần Thơ ngay khi thu thập và tiền xử lý dữ liệu. Chúng tôi có thể nghiên cứu áp dụng cho các vấn đề dự báo tương tự như dự báo mực nước, dự báo lưu lượng cuộc gọi điện thoại,... do các mô hình trong bài là tổng quát cho các vấn đề về dự báo.

TÀI LIỆU THAM KHẢO

1. P. Aksornsingchai, C. Srinilta. Statistical Downscaling for Rainfall and Temperature Prediction in Thailand. *Proc. of the Intl. MultiConference of Engineers and Computer Scientists*, pp. 356-361, (2011).
2. A. Anandhi, V.V. Srinivas, R.S. Nanjundiah, D.N. Kumar. Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine. *International Journal of Climatology*, vol. 28(3):401–420, (2008).
3. A. Bárdossy and E.J. Plate. Space-time model for daily rainfall using atmospheric circulation patterns. *Water Resources Research 28: doi: 10.1029/91WR02589*. ISSN: 0043-1397, (1992).
4. A. Bárdossy. Downscaling from GCMs to local climate through stochastic linkages. *Journal of Environmental Management*, vol. 49(1): 7-17, (1997).
5. T.A. Buishand and T. Brandsma. Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling. *Journal Water Resources Research*, Vol.37(11):2761-2776, (2001).
6. Bộ Tài nguyên và Môi trường. Kịch bản biến đổi khí hậu và nước biển dâng cho Thành phố Cần Thơ. *Báo cáo kỹ thuật, Bộ Tài nguyên và Môi trường, Hà Nội*, (2011).
7. L. Breiman, J.H. Friedman, R.A. Olshen and C. Stone. Classification and Regression Trees. *Wadsworth International*, (1984).
8. L. Breiman. Bagging predictors. *Machine Learning* vol. 24(2):123–140, (1996).
9. L. Breiman. Random forests. *Machine Learning* vol. 45(1):5–32, (2001).
10. C.C. Chang and C.J. Lin. LIBSVM - a library for support vector machines. (2011).
11. S.T. Chen, P.S. Yu, Y.H. Tang. Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. *Journal of Hydrology* 385:13–22, (2010).
12. C.T. Dhanya, D.N. Kumar. Multivariate nonlinear ensemble prediction of daily chaotic rainfall with climate inputs. *Journal of Hydrology, Elsevier*, vol.403(3-4):292-306, (2011).
13. C.T. Dhanya, D.N. Kumar. Data Mining for Evolving Fuzzy Association Rules for Predicting Monsoon Rainfall of India. *Journal of Intelligent Systems, Freund & Pettman, UK*, vol.18(3):193-209, (2010).
14. K. McKuffie and A. Henderson-Sellers, A Climate Modeling Primer, *John Wiley & Sons Ltd., UK*, ISBN 0-470-85750-1, (2005).
15. E. Fix, J. Hodges. Discriminatoiry Analysis: Small Sample Performance. *Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, USA*, (1952).
16. S. Ghosh, P.P. Mujumdar. Statistical downscaling of GCM simulations to

streamflow using relevance vector machine. *Advances in Water Resources*, vol. 31(1):132-146, (2008).

17. S. Ghosh. SVM-PGSL coupled approach for statistical downscaling to predict rainfall from GCM output. *Journal of Geophysical Research: Atmospheres*, vol.115(D22):1984-2012, (2010).
18. M.K. Goyal, C.S.P. Ojha. Evaluation of Various Linear Regression Methods for Downscaling of Mean Monthly Precipitation in Arid Pichola Watershed. *Natural Resources*, vol.1(1):11-18, (2010).
19. M.Z. Hashmi, A.Y. Shamseldin, B.W. Melville. Statistical downscaling of precipitation: state-of-the-art and application of bayesian multi-model approach for uncertainty assessment. *Hydrology and Earth System Sciences Discuss.* (6):6535-6579, (2009).
20. R. Ihaka, R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, vol.5(3):299-314, (1996).
21. J. Murphy. Predictions of climate change over Europe using statistical and dynamical downscaling techniques. *Intl Journal of Climatology*, Vol.20(5):489-501, (2000).
22. J.P. Palutikof, C.M. Goodess, S.J. Watkins and T. Holt. Generating Rainfall and Temperature Scenarios at Multiple Sites: Examples from the Mediterranean. *Journal of Climate*, Vol.15(24): 3529-3548, (2002).
23. A. Pasini. Neural NetworkModeling in Climate Change Studies. *In Artificial Intelligence Methods in the Environmental Sciences, S. E. Haupt et al. (eds.*), pp. 235-254, (2009).
24. J.R. Quinlan. C4.5: Programs for Machine Learning. *Morgan Kaufmann*, (1993).
25. D. Raje, P.P. Mujumdar. A comparison of three methods for downscaling daily precipitation in the Punjab region. *Hydrological Processes*, vol.25(23):3575–3589, (2011).
26. J.S. Selker and D.A. Haith. Development and testing of single-parameter precipitation distributions. *Water Resources Research 26: doi: 10.1029/90WR01648.* ISSN: 0043-1397, (1990).
27. S. Tripathi, V.V. Srinivasa, R.S. Nanjundiahb. Downscaling of precipitation for climate change scenarios: a support vector machine approach. Journal of Hydrology 330:621–640, (2006).
28. C.P. Tung and D.A. Haith. Global-warming effects on New York streamflows. *Journal of Water Resources Planning and Management*, 121(2), pp. 216-225, (1995).
29. V. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, (1995).
30. H. Von Storch, E. Zorita and U. Cubasch. Downscaling of climate change estimates to regional scales: An application to winter rainfall in the Iberian Peninsula. Journal of Climate 6: 1161-1171, (1993).
31. X. Wu and V. Kumar. Top 10 Algorithms in Data Mining. Chapman & Hall/CRC, (2009).
32. H. Yu, S.C. Liu and R.E. Dickinson. Radiative effects of aerosols on the evolution of the atmospheric boundary layer. *Journal of Geophysical Research: Atmospheres*, 107(D12), 4142, doi:10.1029/2001JD000754, (2002).

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

MÔ HÌNH KHẢO SÁT NGƯỠNG ĐÁM ĐÔNG TRỰC TUYẾN VÀ ỨNG DỤNG ĐỀ XUẤT MỘT HỆ THỐNG ĐĂNG KÝ HỌC PHẦN MỚI CHO TRƯỜNG ĐẠI HỌC CẦN THƠ TRÊN NỀN TÍNH TOÁN ĐÁM MÂY VỚI CẤU HÌNH HỢP LÝ

Lê Quyết Thắng[1], Võ Hoàng Tú[2], Mai Yến Trinh[3] và Võ Thị Cẩm Tú[3]

[1] *Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ*
[2] *Trường Cao đẳng Kinh tế - Kỹ Thuật Cần Thơ*
[3] *Trường Đại học Cửu Long*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Assessment survey modelling for the online crowd threshold problem applied to the course registration system in Can Tho University and recommending a new system based on the cloud computing with reasonable configuration

Từ khóa:
Mô hình mạng các hàng chờ, Chỉ số sử dụng, Ngưỡng tắc nghẽn, Thời gian lưu trú, Tính toán đám mây, Máy chủ ảo, Số liệu mô phỏng ngẫu nhiên

Keywords:
Queueing Network Models, Utilisation, Congestion Threshold, Residence Time, Cloud Computing, Virtual Server, Random Simulated Data

ABSTRACT

The congestion problem happens when there is a crowd accessing online has become more frequent. There have been several different solutions to solve this problem, such as upgrading equipments which have capable of faster processing, or using cloud computing with simultaneous virtual server number increasing with the number of online instant crowd. Another aspect of the congestion problem is to forecast itself for having time to prepare response measures, especially for dealing with DDoS attacks by hackers. Based on the queuing network model for a system serving online crowd, we have modeled congestion threshold survey. The application presents results of the survey and the predicted congestion thresholds of online course registration system at Can Tho University. These results allow to propose a new online course registration system based on cloud computing with reasonable configuration.

TÓM TẮT

Bài toán tắc nghẽn khi có đám đông truy cập trực tuyến ngày càng trở nên thường xuyên. Đã có nhiều giải pháp khác nhau để giải quyết vấn đề này, chẳng hạn như nâng cấp thiết bị có khả năng xử lý nhanh hơn nhiều lần, hoặc sử dụng điện toán đám mây với số lượng máy chủ ảo xử lý đồng thời tăng theo số lượng đám đông trực tuyến. Một khía cạnh khác của bài toán tắc nghẽn đó là dự báo nó để có thời gian chuẩn bị các giải pháp ứng phó, đặc biệt khi đối phó với các đợt tấn công DDoS của tin tặc. Dựa trên mô hình mạng các hàng chờ cho một hệ thống phục vụ đám đông trực tuyến, bài báo đã mô hình hóa Phương pháp khảo sát ngưỡng tắc nghẽn. Phần ứng dụng trình bày kết quả khảo sát và dự báo các ngưỡng tắc nghẽn của hệ thống đăng ký học phần trực tuyến tại Trường Đại học Cần Thơ. Kết quả này cho phép đề xuất một hệ thống mới đăng ký học phần dựa trên nền Tính toán đám mây với cấu hình hợp lý.

1 GIỚI THIỆU

Ngày nay việc ứng dụng một dịch vụ trực tuyến trên Internet đã trở nên thông dụng và gần như không thể thiếu đối với các dịch vụ có uy tín. Mặc dù vậy các dịch vụ trực tuyến này luôn phải đối mặt với hai thách thức cơ bản:

– Nhu cầu tham gia trực tuyến tăng nhanh tạo ra đám đông khách hàng có thể làm tắc nghẽn dịch vụ.

– Cạnh tranh không lành mạnh dẫn đến việc sử dụng tin tặc tấn công dưới dạng đám đông "ảo"

làm mất uy tín đối phương, chẳng hạn như tấn công DoS hay DDoS.

Để vượt qua các thách thức này người ta thường hướng tới các công nghệ mới về thiết kế và xử lý các hệ thống xử lý song song hay nâng cấp các thiết bị phần cứng có cấu hình cao hơn nhiều lần. Các giải pháp này rất thụ động vì không xác định được bản chất của hiện tượng tắc nghẽn và dẫn đến hậu quả là mất tiền mà vẫn không giải quyết được tắc nghẽn.

Từ khi có Internet và hiện tượng tắc nghẽn ngày càng trở nên thường xuyên hơn và gây ra nhiều hậu quả về chất lượng phục vụ (QoS), nhiều nhà nghiên cứu đã tập trung phát triển các giải pháp chống tắc nghẽn dựa trên mô hình mạng các hàng chờ.

S. Athuraliya *et al.* (2001) đã đề xuất mô hình Quản trị hàng chờ chủ động (AQM, Active Queue Management) để tính các độ dài hàng chờ tương ứng với hiện tượng tắc nghẽn. Trên cơ sở này nhiều giải pháp khác nhau được xây dựng nhằm chủ động cắt bỏ các gói tin khi độ dài hàng chờ đã đạt tới một độ dài quy định, chẳng hạn như: RED (Random Early Detection) của S. Floyd *et al*. (2001), DRED (Dynamic Random Early Drop) của nhóm J. Aweya *et al*. (2001), Discret Time DRED của AJ Hussein *et al.* (2007). Cơ sở tính toán của các giải pháp dựa trên AQM là sử dụng bộ đếm các gói tin truy cập tại router để quản trị các hàng chờ. Như vậy muốn ứng dụng AQM một cách tương tự cho các hệ thống trực tuyến trên Internet cần phải cài đặt bộ đếm các gói tin.

Vấn đề hiện nay là rất nhiều hệ thống xử lý trực tuyến không có các bộ đếm thời gian thực các truy cập hệ thống nên không thể sử dụng các mô hình AQM. Bài báo này vẫn tiếp cận mô hình mạng các hàng chờ để xây dựng một giải pháp đơn giản nhằm dự báo hiện tượng tắc nghẽn đối với các hệ thống phục vụ đám đông. Kết quả dự báo có thể hỗ trợ tốt cho các giải pháp phòng ngừa tắc nghẽn.

Phương pháp này được ứng dụng vào khảo sát Hệ thống đăng ký học phần trực tuyến tại Trường Đại học Cần Thơ. Kết quả khảo sát này được ứng dụng tiếp để khảo sát hệ thống mới dựa trên nền Tính toán đám mây với khả năng tiết kiệm tài nguyên cao.

2 CÁC THAM SỐ ĐÁNH GIÁ MẠNG CÁC HÀNG CHỜ

2.1 Các tham số cơ bản

Mạng các hàng chờ (Queueing Network) là một phương pháp mô hình hóa một hệ thống nối mạng các hàng chờ (xem Edward D. Lazowska *et al*. (1984)). Kiến trúc tổng quát của một mạng các hàng chờ gồm các thành phần cơ bản: trung tâm Phục vụ (Service Center), trung tâm Trì hoãn (Delay Center) và Luồng tải công việc (Workload Flow). Trung tâm Phục vụ tuân theo mô hình xếp hàng. Trung tâm Trì hoãn là một mô hình của các thời gian chậm trễ trên mạng, như thời gian truyền dữ liệu hoặc suy nghĩ của khách hàng trước khi yêu cầu hệ thống xử lý. Luồng tải công việc kết nối các trung tâm và tuân theo một số luật trên mạng các hàng chờ, như luật Ép luồng (Forced Flow Law) và luật cân bằng tải (Balanced Flow Law) (Lê Quyết Thắng và Phạm Nguyên Khang, 2013).

Mô hình xếp hàng áp dụng cho mỗi trung tâm phục vụ trong trường hợp xử lý đám đông có thể xấp xỉ bởi mô hình M/M/1. Giả thiết này là hợp lý.

Thứ nhất, những khách hàng có nhu cầu khác đám đông sẽ có mật độ thấp hơn rất nhiều, do đó tỷ lệ chia sẻ thời gian xử lý cho nhóm khách hàng này sẽ không đáng kể. Như vậy có thể sử dụng một lớp khách hàng có cùng nhu cầu và tạo ra đám đông cần được xử lý. Khi đó mô hình xếp hàng tại mỗi trung tâm phục vụ là M/M/s.

Thứ hai, hầu hết các hệ thống ứng dụng trực tuyến đều đơn giản hóa quá trình phân tải. Có nghĩa là mỗi hệ thống xử lý trực tuyến đều cài sẵn module phân tải cho s máy chủ xử lý song song khi có đám đông xảy ra. Với phương pháp phân tải như vậy, s máy chủ xử lý song song sẽ tạo ra s trung tâm phục vụ song song. Khi đó mô hình xếp hàng tại mỗi trung tâm phục vụ sẽ trở thành M/M/1.

Xét các tham số trên mạng các hàng chờ và các công thức cơ bản:

N – Số khách hàng đăng nhập hệ thống và được lưu giữ tại một trung tâm trì hoãn trước khi tham gia một dịch vụ nào đó có trong hệ thống.

p – Xác suất để một khách hàng đã đăng nhập đi tới một trung tâm phục vụ.

d – Mật độ khách hàng truy cập hệ thống.

λ – Lưu lượng giao dịch trung bình của luồng đi tới một trung tâm phục vụ. Dưới mô hình xếp hàng λ còn được hiểu là mật độ trung bình khách hàng có nhu cầu được phục vụ. Từ ý nghĩa của N, p và λ, ta có quan hệ đầu tiên: $\lambda = pN$.

R – Thời gian lưu trú (Residence Time) trung bình của mỗi khách hàng tại một trung tâm phục vụ. Dưới mô hình xếp hàng thì R còn được gọi là thời gian đáp ứng (Response Time) một yêu cầu.

R* - Thời gian trung bình hoàn tất tất cả các xử lý tại tất cả các trung tâm phục vụ của một khách hàng (xem Lazowska (1984)) và được tính theo công thức:

$R* = \sum_i R_i$, i chạy trên tất cả các trung tâm phục vụ.

Q – Độ dài hàng (Queue Length) trung bình trong một trung tâm phục vụ, hay còn là số khách hàng trung bình đang lưu trú tại trung tâm. Mặt khác, theo luật Little (xem Lê Quyết Thắng và Phạm Nguyên Khang (2013)), ta lại có quan hệ: $Q = \lambda R$.

S – Thời gian phục vụ (Service Time) trung bình một yêu cầu của khách hàng tại một trung tâm phục vụ.

φ – Chỉ số sử dụng (Utilisation) dịch vụ của một trung tâm phục vụ, hay còn được hiểu là mức tải của trung tâm.

Từ mô hình M/M/1 của Lý thuyết xếp hàng ta có công thức tính chỉ số sử dụng: $\varphi = \lambda S$ (xem Lê Quyết Thắng và Phạm Nguyên Khang (2013)). Nhưng $\lambda = pNS$, nên $\varphi = pNS$.

2.2 Chỉ số tắc nghẽn

Điều kiện để hệ thống không bị tắc nghẽn (xem Lê Quyết Thắng và Phạm Nguyên Khang (2013)) là: $0 < \varphi < 1$. Nhưng trong thực tế, dấu hiệu tắc nghẽn xảy ra khi $\varphi > 0.7$, do thời gian lưu trú R trong một trung tâm phục vụ sẽ tăng lên rất nhanh và hướng tới vô cùng.

Thật vậy, thời gian lưu trú tại một trung tâm phục vụ với mô hình M/M/1 có thể được tính theo chỉ số sử dụng (xem Lê Quyết Thắng và Phạm Nguyên Khang (2013)) như sau:

Trước hết từ M/M/1 ta có công thức tính $Q = \varphi/(1-\varphi)$. Do $Q = \lambda R$ và $\varphi = \lambda S$, nên:

$$R = S/(1-\varphi) \quad (1)$$

Như vậy có thể xem công thức tính R là một hàm số dạng Hyperbol biến thiên theo φ và có vận tốc biến thiên là đạo hàm bậc 1 của R theo φ, hay $R' = S/(1-\varphi)^2$.

Nếu khảo sát vận tốc tăng R' của thời gian lưu trú theo φ có thể thấy các đặc điểm cơ bản:

– Khởi động $\varphi = 0$, thì vận tốc tăng là S.

– Khi φ đạt 0.5, thì vận tốc tăng lên: 4S.

– Cho φ tăng lên tới 0.7 thì vận tốc tăng nhanh và vượt quá 10.5S.

– Cho φ tăng tới 0.8 vận tốc tăng rất nhanh và vượt 25S.

– Cho φ vượt 0.9 vận tốc tăng hơn 100S và hướng tới vô cùng.

Kết quả khảo sát nhanh này cho thấy sự chậm chạp của hệ thống sẽ xảy ra khi có một trung tâm phục vụ trong hệ thống có $\varphi > 0.7$. Vì vậy φ có thể được gọi là *chỉ số tắc nghẽn* và "sự chậm chạp" còn có thể chấp nhận được tương ứng với φ nằm trong khoảng [0.7, 0.8].

3 KHẢO SÁT HỆ THỐNG ĐĂNG KÝ HỌC PHẦN TRỰC TUYẾN TẠI TRƯỜNG ĐẠI HỌC CẦN THƠ

3.1 Xây dựng mô hình mạng các hàng chờ

3.1.1 Mô tả tình huống

Trong chương trình đào tạo theo tín chỉ tại Trường Đại học Cần Thơ, sinh viên phải đăng ký môn học trực tuyến trên một hệ thống thông tin theo quy trình sau:

– Tất cả sinh viên đều phải mở tài khoản của mình trong hệ thống để có thể đăng ký môn học trực tuyến.

– Khi đăng ký môn học, mỗi sinh viên đều phải đăng nhập. Để hệ thống có thể chấp nhận số lượng N sinh viên đăng ký môn học trực tuyến đồng thời đủ lớn, người ta bố trí hai máy chủ Đăng ký học phần và một máy chủ Phân tải. Máy chủ Phân tải có nhiệm vụ cân bằng tải cho hai máy chủ Đăng ký học phần, do đó nó sẽ kiểm soát và điều hướng các đăng nhập sang máy chủ nào có số lượng đăng ký trực tuyến ít hơn.

– Khi thực hiện đăng ký học phần, mỗi sinh viên phải suy nghĩ và điền vào mẫu đăng ký học phần (trực tuyến). Sau khi hoàn tất mẫu, sinh viên gửi kết quả cho máy chủ Đăng ký học phần. Máy chủ Đăng ký học phần xử lý mẫu và chuyển kết quả xử lý cho máy chủ Cơ sở dữ liệu (CSDL) để lưu lại các đăng ký mới của sinh viên.

– Sau mỗi lần đăng ký thành công sinh viên có thể xem lại kết quả và thoát khỏi hệ thống.

3.1.2 Phân tích các trung tâm trì hoãn và trung tâm phục vụ

Máy chủ đăng nhập xử lý các đăng nhập của sinh viên sau đó phân tải tạo thành trung tâm phục vụ đầu tiên ở mức 0. Ta gọi trung tâm này là trung tâm 0 (Đăng nhập và Phân tải).

Sinh viên đăng nhập thành công và được phân tải sang máy chủ Đăng ký 1 hoặc 2. Trước khi

đăng ký sinh viên phải cần thời gian suy nghĩ và điền đầy đủ mẫu đăng ký học phần trực tuyến. Như vậy, Máy chủ Đăng ký học phần 1 hoặc 2 phải tạo ra vùng nhớ riêng để lưu lại các sinh viên đăng nhập và cần thời gian suy nghĩ. Các vùng nhớ này tạo thành các trung tâm Trì hoãn. Tất cả sinh viên đăng nhập thành công đều được xếp vào một trong hai trung tâm Trì hoãn ở mức 1 để chuẩn bị đăng ký học phần. Máy chủ Đăng ký 1 tạo ra trung tâm Trì hoãn 1.1, còn máy chủ Đăng ký 2 tạo ra trung tâm Trì hoãn 1.2.

Sau khi hoàn thành mẫu đăng ký học phần, sinh viên yêu cầu được đăng ký chính thức. Máy chủ Đăng ký học phần sẽ thực hiện công đoạn xử lý mẫu đăng ký và chuyển cho máy chủ CSDL cập nhật: thêm, sửa hoặc xóa các đăng ký học phần của sinh viên. Với chức năng này, các máy chủ Đăng ký học phần sẽ tạo ra các trung tâm phục vụ ở mức 2. Ta gọi trung tâm Phục vụ tương ứng với Máy chủ Đăng ký 1 là trung tâm Phục vụ 2.1 (Đăng ký học phần 1), còn trung tâm tương ứng với Máy chủ đăng ký 2 là trung tâm Phục vụ 2.2 (Đăng ký học phần 2).

Máy chủ CSDL cập nhật các dữ liệu đăng ký bởi sinh viên và tạo ra trung tâm Phục vụ ở mức 3 và được gọi là trung tâm Phục vụ 3 (CSDL).

3.1.3 Phân tích luồng

Do mục tiêu đánh giá khả năng tắc nghẽn trong hệ thống, nên chúng ta chỉ quan tâm đến các luồng lớn như phần phân tích dưới đây.

Luồng sinh viên đăng nhập đi vào trung tâm 0 là luồng khởi tạo sự hoạt động của hệ thống và có lưu lượng là λ.

Thoát khỏi trung tâm 0 với luồng λ không đổi, sinh viên được gửi đến trung tâm 1.1 với xác suất là q_1 hoặc đến trung tâm 1.2 với xác suất tương ứng là q_2. Do cân bằng tải nên $q_1 = q_2 = q = 1/2$.

Tại trung tâm 1.1 hoặc 1.2, sinh viên có thể suy nghĩ để điều chỉnh Kế hoạch học tập, sau đó điền mẫu Đăng ký học phần mới của mình.

Từ trung tâm 1.1, mỗi sinh viên được chuyển đến trung tâm 2.1 với xác xuất p_1. Tương tự, từ trung tâm 1.2, sinh viên được chuyển đến trung tâm 2.2 với xác suất p_2. Do sinh viên có khả năng điều chỉnh Kế hoạch học tập và điền mẫu Đăng ký học tập như nhau nên họ có cùng xác suất đi tới trung tâm Đăng ký học phần, vậy: $p_1 = p_2 = p$.

Thoát khỏi trung tâm Đăng ký học phần, mỗi sinh viên sẽ có 3 trạng thái:

Trạng thái 1: Đăng ký mới tạo luồng đi tới trung tâm CSDL,

Trạng thái 2: Kết thúc cập nhật tạo luồng quay về trung tâm Trì hoãn và

Trạng thái 3: Hoàn tất đăng ký học phần tạo luồng thoát khỏi hệ thống.

Do các sinh viên đều thực hiện quy trình đăng ký học phần tương tự nhau nên mỗi sinh viên dù thoát khỏi trung tâm 2.1 hay 2.2 đều có phân phối xác suất giống nhau theo 3 trạng thái trên. Như vậy mỗi sinh viên khi thoát khỏi trung tâm 2.1 hoặc 2.2 đều có phân phối xác suất:

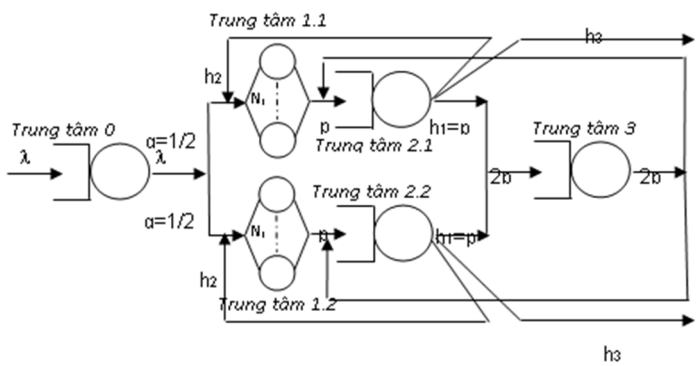
P(Trạng thái 1) = h_1, P(Trạng thái 2) = h_2 và P(Trạng thái 3) = h_3.

Dễ dàng nhận thấy, mỗi sinh viên hoàn thành đăng ký học phần mới, bắt buộc sẽ cập nhật CSDL, do đó: $h_1 = p$.

Ở mức 3, mỗi sinh viên hoàn thành cập nhật đăng ký học phần của mình sẽ tạo một lưu lượng tương ứng với đầu vào là 2p để quay về mức 2.

3.1.4 Mô hình mạng các hàng chờ

Kết quả phân tích trung tâm và luồng được tổng hợp thành mô hình Mạng các hàng chờ của hệ thống Đăng ký học phần trực tuyến tại Trường Đại học Cần Thơ và được minh họa trong Hình 1.



Hình 1: Mô hình Mạng các hàng chờ của Hệ thống đăng ký học phần trực tuyến tại Trường Đại học Cần Thơ

3.2 Tính các tham số ổn định không phụ thuộc lưu lượng luồng

3.2.1 Hệ số kinh nghiệm

Do chỉ khảo sát chỉ số tải khi có đám đông đăng ký học phần nên ở đây ta chỉ quan tâm đến các tham số ổn định liên quan đến các trung tâm 2.1, 2.2 và 3.

Các tham số ổn định được quan tâm gồm:

Xác suất p (xác suất để mỗi sinh viên hoàn thành điều chỉnh Kế hoạch học tập và điền mẫu đăng ký học phần để đi vào trung tâm Đăng ký học phần);

Thời gian S xử lý một mẫu đăng ký tại trung tâm 2.1 hoặc 2.2 và thời gian S_{CSDL} cập nhật Đăng ký học phần tại trung tâm 3.

Đối với hệ thống đăng ký học phần tại Trường Đại học Cần Thơ, chúng tôi khảo sát kinh nghiệm của các nhà quản lý hệ thống và chỉ nhận được 3 dữ liệu trung bình sau:

– Mỗi máy chủ Đăng ký học phần chạy chậm khi có khoảng N1 = 400 sinh viên đăng nhập đồng thời.

– Máy chủ CSDL chạy chậm khi có tổng sinh viên đăng nhập đồng thời khoảng N = 500.

– Thời gian trung bình mà hệ thống hoàn tất xử lý và cập nhật một mẫu đăng ký học phần dao động từ 2 - 4 giây. Như vậy thời gian trung bình là khoảng 4 giây khi toàn bộ hệ thống đạt ngưỡng đăng nhập khoảng N = 500 sinh viên.

Để tính các tham số ổn định p, S và S_{CSDL}, cần xây dựng hệ ba phương trình tìm các tham số này. Mặt khác như đã khảo sát ở phần trên: "sự chậm chạp" của hệ thống tương ứng với "sự chậm chạp" tại một trung tâm phục vụ nào đó và ở đó sẽ có chỉ số sử dụng $\varphi \in [0.7, 0.8]$. Nhưng ta lại cần một hệ số tương ứng với sự chậm chạp mà không phụ thuộc vào một trung tâm phục vụ nào. Ta đặt: $k \in [0.7, 0.8]$ là một hệ số chỉ "sự chậm chạp" của hệ thống. Như vậy, nếu xét "sự chậm chạp" xảy ra tại một trung tâm Đăng ký học phần thì $\varphi = k$, còn nếu xảy ra tại trung tâm CSDL thì $\varphi_{CSDL} = k$.

3.2.2 Xây dựng Phương trình 1

Giả sử số sinh viên đăng nhập đồng thời tại một máy chủ Đăng ký học phần đạt mức tối đa: N_1 = 400 sinh viên và dẫn đến hậu quả máy chủ Đăng ký học phần chạy chậm. Khi đó chỉ số tải tương ứng với dấu hiệu chậm chạp tại máy chủ Đăng ký này là $\varphi = k$. Nếu đặt λ_1 là lưu lượng sinh viên đi tới một trung tâm mức 2, khi đó $\lambda_1 = pN_1$. Như vậy tại trung tâm 2.1 (hoặc 2.2) chỉ số tải: $\varphi = \lambda_1 S = pSN_1 = k$. Ta nhận được phương trình thứ nhất:

$$N_1 pS = k \quad (2)$$

3.2.3 Xây dựng Phương trình 2

Tương tự đối với trung tâm 3, ta sẽ xét trường hợp có N = 500 đăng ký học phần cần cập nhật và được phát sinh từ hai trung tâm 2.1 và 2.2. Khi đó có N'_1=250 sinh viên đăng nhập đồng thời tại mỗi trung tâm 2.1 và 2.2. Mặt khác lưu lượng truy cập CSDL xuất phát từ hai trung tâm 2.1 và 2.2 và được bảo toàn theo luật Ép luồng, do đó lưu lượng này bằng: $\lambda_{CSDL} = pN'_1 + pN'_1 = 2pN'_1 = pN$.

Do: $\varphi_{CSDL} = \lambda_{CSDL} S_{CSDL} = pNS_{CSDL}$ và suy ra phương trình thứ 2:

$$NpS_{CSDL} = k \quad (3)$$

3.2.4 Xây dựng Phương trình 3

Để tính R*, ta sử dụng tổng số sinh viên đăng nhập là N = 500 kết nối các trung tâm 2.1, 2.2 và 3, khi đó sẽ có số N'_1 = 250 sinh viên đăng nhập đồng thời lưu trú tại mỗi trung tâm Trì hoãn 1.1 và 1.2.

Bây giờ sử dụng (2) cho trung tâm 2.1 (hoặc trung tâm 2.2) khi N'_1 = 250 và $\varphi = \lambda_1 S = pSN'_1$ dẫn đến:

$$R = \frac{S}{1 - pSN'_1}, \text{ hay: } R = \frac{S}{1 - 250pS} \quad (4)$$

Còn áp dụng cho trung tâm 3 với N = 500, ta có tiếp:

$$R_{CSDL} = \frac{S_{CSDL}}{1 - pS_{CSDL}N}, \text{ hay:}$$

$$R_{CSDL} = \frac{S_{CSDL}}{1 - 500pS_{CSDL}} \quad (5)$$

Do:

$$R^* = 2R + R_{CSDL} = 4 \text{ (s)}$$

Nên từ (4) và (5) ta tính được:

$$R^* = \frac{2S}{1 - 250pS} + \frac{S_{CSDL}}{1 - 500pS_{CSDL}} = 4 \text{ (s).}$$

Ta nhận được phương trình thứ 3:

$$\frac{2S}{1 - 250pS} + \frac{S_{CSDL}}{1 - 500pS_{CSDL}} = 4 \quad (6)$$

3.2.5 Giải hệ phương trình

Tổng hợp các kết quả ta có hệ ba phương trình (2), (3) và (6) với N_1= 400 và N = 500. Sử dụng phương pháp thế để giải hệ phương trình này và có kết quả:

$$p = \frac{k}{400S} \quad (7)$$

$$S_{CSDL} = \frac{4S}{5} \quad (8)$$

$$S = \frac{5(8-5k)(1-k)}{28-25k} \quad (9)$$

Tùy trường hợp của k mà tính được S từ (9), sau đó tính p và S_{CSDL} từ (7) và (8).

Nếu k = 0.7, ta tính được: S = 0.643 (s), p = 0.0027 và S_{CSDL} = 0.514 (s).

Còn nếu k = 0.8 thì: S = 0.5 (s), p = 0.004 và S_{CSDL} = 0.4 (s).

Do hệ số kinh nghiệm k là một chỉ số tương đối nên các tham số ổn định p, S và S_{CSDL}cũng là tương đối và được ước lượng trong các khoảng biến động hợp lý. Như vậy, nếu k ∈ [0.7, 0.8] thì p ∈ [0.0027, 0.004], S ∈ [0.5, 0.643] và S_{CSDL} ∈ [0.4, 0.514].

Với các tham số này chúng ta có thể khảo sát các chỉ số tắc nghẽn của hệ thống vật lý để đưa ra các ngưỡng dự báo hiện tượng tắc nghẽn. Nhưng mục tiêu của bài báo là khảo sát hệ thống mới hợp lý hơn trên nền Tính toán đám mây. Trong khi đó chỉ có tham số p không phụ thuộc vào kiến trúc của hệ thống, do đó chỉ có p được sử dụng để khảo sát hệ thống mới này.

4 KHẢO SÁT HỆ THỐNG ĐĂNG KÝ HỌC PHẦN TRỰC TUYẾN TRÊN NỀN TÍNH TOÁN ĐÁM MÂY

4.1 Mô hình Tính toán đám mây

Trong mô hình của hệ thống Đăng ký học phần người ta sử dụng 4 máy chủ vật lý, trong đó có 2 máy chủ sử dụng riêng cho xử lý Đăng ký học phần của sinh viên. Đây là cấu hình được sử dụng khi chưa có kế hoạch hợp lý, nhưng khi đã có kế hoạch hợp lý thì cấu hình trên đã tỏ ra dư thừa tài nguyên. Mô hình mới dựa trên nền Tính toán đám mây cho phép quy hoạch tài nguyên hợp lý hơn.

Quy hoạch tài nguyên hợp lý nhằm vào quy hoạch máy chủ xử lý Đăng ký học phần. Tùy vào mật độ sinh viên đăng ký học phần mà ta sử dụng số lượng máy chủ ảo Đăng ký học phần từ 1 đến số tối đa cho phép. Như vậy nếu mật độ sinh viên truy cập không lớn thì phần tài nguyên dành cho các máy chủ ảo không dùng tới có thể sử dụng cho các công việc khác.

4.1.1 Cấu hình máy chủ ảo

Do công nghệ Tính toán đám mây chỉ có thể cung cấp cấu hình máy chủ ảo theo những phương án xác định trước, nên khi đưa ra mô hình xử lý trực tuyến trên nền Tính toán đám mây chúng ta cũng chỉ có thể xác định trước một số cấu hình cho máy chủ ảo. Do cấu hình máy chủ ảo phụ thuộc vào máy chủ vật lý, nên cấu hình máy chủ ảo chỉ mang tính tương đối.

Giả thiết rằng các máy chủ vật lý dành riêng cho hệ thống Đăng ký học phần trực tuyến của Trường Đại học Cần Thơ được sử dụng để cung cấp các máy chủ ảo. Khi đó ta có thể đề nghị cung cấp ba mức cấu hình: (1) Cấu hình Mạnh: máy chủ vật lý có thể sinh tối đa 4 máy chủ ảo, (2) Cấu hình Trung: có thể sinh tối đa 7 máy chủ ảo và (3) Cấu hình Yếu: có thể sinh tối đa 14 máy chủ ảo.

a. Cấu hình Mạnh

Trong mô hình mạng các hàng chờ của hệ thống Đăng ký học phần các máy chủ Đăng nhập và CSDL được sử dụng chung cho một hoặc nhiều máy chủ Đăng ký học phần. Vì vậy, chúng ta có thể sử dụng cấu hình Mạnh cho Máy chủ Đăng nhập và máy chủ CSDL.

Mục tiêu của bài báo là khảo sát khả năng quy hoạch tài nguyên cho máy chủ Đăng ký học phần nên các thông số cho máy chủ Đăng nhập sẽ không bàn tới và chỉ xét một máy chủ ảo với cấu hình Mạnh cho máy chủ Đăng nhập.

Do cấu hình Mạnh có số tối đa được sinh ra là 4 từ máy chủ vật lý, nên nếu sử dụng cấu hình Mạnh cho máy chủ ảo CSDL thì số truy cập CSDL dẫn đến "sự chậm chạp" của hệ thống sẽ là N^*_{CSDL} = 500/4 = 125. Trường hợp sử dụng cấu hình Mạnh cho máy chủ ảo xử lý Đăng ký học phần, số truy cập dẫn đến "sự chậm chạp" sẽ là N^*_M = 400/4 = 100.

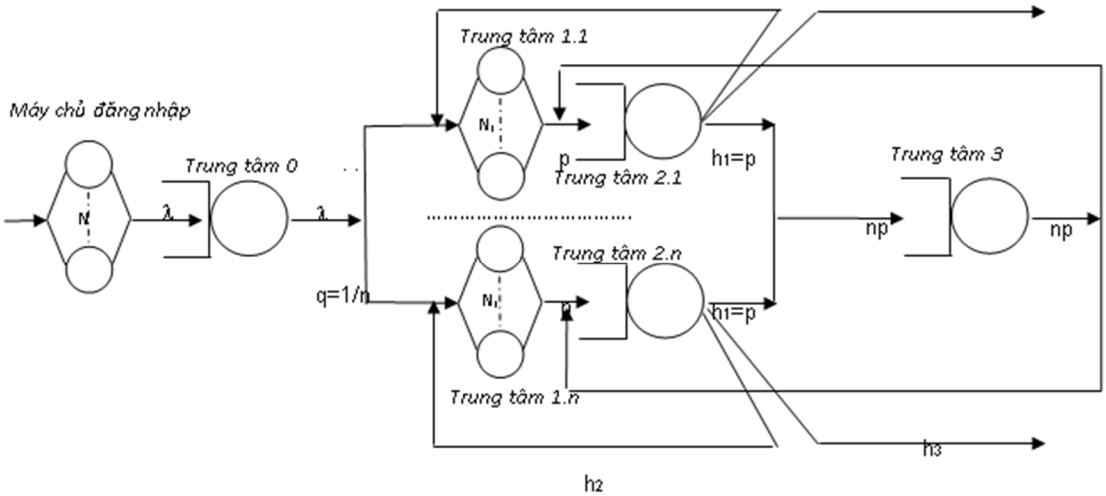
b. Cấu hình Trung

Cấu hình này chỉ sử dụng cho máy chủ ảo xử lý Đăng ký học phần khi có mật độ truy cập của sinh viên không lớn.

Nếu máy chủ vật lý sinh tối đa là 7 máy chủ ảo thì số lượng truy cập dẫn đến "sự chậm chạp" sẽ khoảng N^*_T = 55 (≈ 400/7).

c. Cấu hình Yếu

Tương tự như cấu hình Trung, nếu máy chủ vật lý sinh tối đa là 14 máy ảo thì số lượng truy cập dẫn đến "sự chậm chạp" khoảng $N^*_Y = 25$ (≈ 400/14).



Hình 2: Mô hình Mạng các hàng chờ của Hệ thống đăng ký học phần trực tuyến trên nền Tính toán đám

4.1.2 Mô hình trên nền Tính toán đám mây

Tương tự như phân tích các trung tâm và luồng đối với hệ thống vật lý, chúng ta xây dựng được mô hình mạng các hàng chờ cho hệ thống Đăng ký học phần trực tuyến trên nền Tính toán đám mây như trong Hình 2.

Các trung tâm 1.1 và 2.1 là máy chủ ảo xử lý học phần thứ nhất khi mật độ truy cập thấp. Khi mật độ truy cập tăng lên thì số lượng máy chủ ảo xử lý Đăng ký học phần cũng tăng lên cho phù hợp để tránh tắc nghẽn. Như vậy nếu máy chủ ảo tăng đên số n thì ta có các trung tâm tương ứng là trung tâm 1.n và 2.n.

Các trung tâm 0 (Đăng nhập) và 3 (CSDL) sẽ tùy tình huống mà ta sử dụng máy chủ vật lý hay ảo.

4.2 Khảo sát ngưỡng tắc nghẽn cho mô hình trên nền Tính toán đám mây

Do bản chất của xác suất p, xuất hiện yêu cầu xử lý Đăng ký học phần của một sinh viên, là ổn định và chỉ phụ thuộc vào môi trường đào tạo tại Trường Đại học Cần Thơ, nên tùy hệ số kinh nghiệm k mà ta có các giá trị p tương ứng đã được tính và có thể ước lượng trong khoảng [0.0027, 0.004].

Từ Hình 2 ta thấy toàn bộ luồng lưu thông trên hệ thống tập trung vào trung tâm 0 và trung tâm 3. Do trung tâm 3 là CSDL nên khả năng tắc nghẽn lớn nhất nằm ở đây. Việc quy hoạch các máy chủ Đăng ký học phần ở các trung tâm 1.1, 2.1 đến 1.n và 2.n sẽ phụ thuộc vào trạng thái tắc nghẽn hay không ở trung tâm 3. Vì vậy, ta bắt đầu từ việc xét các ngưỡng tắc nghẽn tại trung tâm 3 sau đó sẽ suy ra số lượng n các máy chủ Đăng ký học phần hợp lý.

4.2.1 Ngưỡng tắc nghẽn đối với trung tâm 3

a. Trường hợp máy chủ CSDL là máy ảo với cấu hình Mạnh: $N = N^*_{CSDL} = 125$

Từ phương trình (3) với k = 0.7, ta tính được $\text{Min}(S^*_{CSDL}) = 1.4$ (s). Còn với k = 0.8 ta tính được: $\text{Max}(S^*_{CSDL}) = 2.351$ (s).

Sử dụng công thức $\varphi^*_{CSDL} = \lambda S^*_{CSDL}$ và chỉ số tắc nghẽn $\varphi^*_{CSDL} = 0.7$, ta tính được ngưỡng dự báo mật độ truy cập dẫn đến tắc nghẽn tại CSDL:

$\text{Min}(\lambda) = 0.7/\text{Max}(S^*_{CSDL}) = 0.298$ (sv/s) và

$\text{Max}(\lambda) = 0.8/\text{Min}(S^*_{CSDL}) = 0.571$ (sv/s).

b. Trường hợp máy chủ CSDL là máy chủ vật lý với N = 500

Một cách tương tự như trên ta lại tính được:

$\text{Min}(S^*_{CSDL}) = 0.4$ và $\text{Max}(S^*_{CSDL}) = 0.514$.

$\text{Min}(\lambda) = 0.7/\text{Max}(S^*_{CSDL}) = 1.361$ và

$\text{Max}(\lambda) = 0.8/\text{Min}(S^*_{CSDL}) = 2$.

4.2.2 Ngưỡng tắc nghẽn đối với các trung tâm 2.i

a. Trường hợp trung tâm 3 và các trung tâm 2.i đều là máy chủ ảo với cấu hình Mạnh

Ngưỡng dự báo tắc nghẽn lúc này là Min(λ)= 0.298 và Max(λ) = 0.571, trong khi máy chủ ảo của trung tâm 2.i có cấu hình Mạnh với $N^*_M = 100$.

Từ phương trình (2): $N_1pS = k$ với k =0.7, suy ra: Min(S^*) = 0.7/(100*0.004) = 1.75, còn với k = 0.8, ta lại tính được Max(S^*) = 0.8/(100*0.0027) = 2.963.

Nếu n là số lượng máy chủ ảo tại các trung tâm 2.i, chỉ số ứng dụng tại một trung tâm sẽ là:

$\varphi^*_1 = (\lambda/n)S^*$. Từ đây suy ra các ngưỡng cho trung tâm 2.1.

Khi λ đạt ngưỡng dự báo tắc nghẽn Min(λ) và φ^*_1 = 0.8 thì Min(n) = [Min(λ)Max(S^*)/0.8] =1.

Khi λ đạt ngưỡng tắc nghẽn Max(λ) và φ^*_1 = 0.7 thì Max(n) = [Max(λ)Max(S^*)/0.7] =2.

Vậy nếu sử dụng máy chủ CSDL là máy ảo với cấu hình Mạnh thì số lượng n máy chủ ảo với cấu hình Mạnh cho xử lý Đăng ký học phần dao động trong khoảng từ 1 đến 2 và chỉ đáp ứng mật độ truy cập tối đa Max(λ) = 0.571.

b. Trường hợp trung tâm 3 là máy ảo có cấu hình Mạnh còn các trung tâm 2.i đều là máy chủ ảo với cấu hình Trung

Tương tự như trên từ phương trình (2) với N^*_T = 55, ta tính được Max(S^*) = 0.8/(55*0.0027) = 5.387 và Min(S^*) = 3.182.

Khi λ đạt ngưỡng dự báo tắc nghẽn Min(λ) và φ^*_1 = 0.8 thì Min(n) = [Min(λ)Max(S^*)/0.8] =1.

Khi λ đạt ngưỡng tắc nghẽn Max(λ) và φ^*_1 = 0.7 thì Max(n) = [Max(λ)Max(S^*)/0.7] = 4.

Vậy nếu sử dụng máy chủ CSDL là máy ảo với cấu hình Mạnh thì số lượng n máy chủ ảo với cấu hình Trung cho xử lý Đăng ký học phần dao động trong khoảng từ 1 đến 4 và chỉ đáp ứng mật độ truy cập tối đa Max(λ) = 0.571.

c. Trường hợp trung tâm 3 là máy ảo có cấu hình Mạnh còn các trung tâm 2.i đều là máy chủ ảo với cấu hình Yếu

Tương tự các kết quả lần lượt là: Max(S^*) = 0.8/55*0.0027= 11.755 và Min(S^*) = 7. Suy ra số lượng máy chủ ảo tương ứng: Min(n) = 3, Max(n) = 9.

Vậy nếu sử dụng máy chủ CSDL là máy ảo với cấu hình Mạnh thì số lượng n máy chủ ảo với cấu hình Yếu cho xử lý Đăng ký học phần dao động trong khoảng từ 1 đến 3 hoặc 9 và chỉ đáp ứng mật độ truy cập tối đa Max(λ) = 0.571.

a) Trường hợp trung tâm 3 là máy chủ vật lý và các trung tâm 2.i đều là máy chủ ảo với cấu hình Mạnh.

Ta cũng vẫn có các phép tính tương tự và nhận được: Min(n) = 3, Max(n) = 8.

Vậy nếu sử dụng máy chủ CSDL là máy vật lý thì số lượng n máy chủ ảo với cấu hình Mạnh cho xử lý Đăng ký học phần dao động trong khoảng từ 1 đến 3 hoặc 8 và có thể đáp ứng mật độ truy cập tối đa Max(λ) = 2.

4.3 Quy hoạch hệ thống trên nền Tính toán đám mây

Kết quả xử lý số liệu sinh viên Đăng ký học phần năm 2012 tại Trường Đại học Cần Thơ cho thấy mật độ sinh viên truy cập đăng ký học phần cao nhất là 0.5 sv/s. Sau đó phần lớn các khoảng thời gian khác mật độ truy cập nhỏ hơn 0.2.

Ta chia thời gian hoạt động của hệ thống Đăng ký học phần thành ba giai đoạn: (1) Giai đoạn Cao điểm có mật độ sinh viên truy cập từ đến 0.5 sv/s, (2) Giai đoạn Trung điểm có mật độ sinh viên truy cập đến 0.2 sv/s và (3) Giai đoạn Bình thường có mật độ sinh viên truy cập đến 0.1 sv/s.

Dựa vào kết quả tính toán ở mục 4.2 và mật độ truy cập thực tế của sinh viên hiện nay sau khi có kế hoạch Đăng ký học phần theo khoa chuyên ngành, toàn bộ hệ thống Đăng ký học phần trực tuyến có thể đưa lên đám mây với cấu hình tùy thuộc vào giai đoạn.

− Giai đoạn Cao điểm: sử dụng cấu hình Mạnh cho tất các các máy chủ.

− Giai đoạn Trung điểm: sử dụng cấu hình Mạnh cho Máy chủ Đăng nhập và CSDL, còn cấu hình Trung dành cho máy chủ Đăng ký học phần.

− Giai đoạn Thấp điểm: sử dụng cấu hình Mạnh cho Máy chủ Đăng nhập và CSDL, còn cấu hình Thấp dành cho máy chủ Đăng ký học phần.

5 KẾT LUẬN

Việc nghiên cứu mô hình mạng các hàng chờ cho một hệ thống ứng dụng trực tuyến cho phép đánh giá được bản chất của hiện tượng tắc nghẽn và do đó có thể đưa ra nhiều mô hình hiệu quả theo hai xu hướng:

(1) Dự báo tắc nghẽn theo mật độ truy cập để tiếp cận các giải pháp chống tắc nghẽn kịp thời. Đặc biệt hiệu quả khi đưa ra các mô hình lọc bớt những khách hàng "không mong muốn" bị nghi ngờ tham gia tấn công DDoS.

(2) Xây dựng mô hình Tính toán đám mây cho cùng một hệ thống nhưng đặc biệt hiệu quả cho tiết kiệm tài nguyên.

Kết quả của bài báo đi theo xu hướng thứ hai và ứng dụng đề xuất mô hình Tính toán đám mây cho hệ thống Đăng ký học phần trực tuyến của Trường Đại học Cần Thơ hiện nay.

Xu hướng thứ hai còn có thể ứng dụng khảo sát những hệ thống trực tuyến để xây dựng mô hình Tính toán đám mây thay thế an toàn hơn và tiết kiệm hơn dựa trên nền công nghệ phần cứng phù hợp.

LỜI CẢM TẠ

Nhóm tác giả chân thành cám ơn anh Lê Hữu Bình, Quản trị viên Hệ thống Đăng ký học phần Trường Đại học Cần Thơ, đã tận tình cung cấp thông tin cho bài báo này.

TÀI LIỆU THAM KHẢO

1. Edward D. Lazowska, John Zahorjan,G. Scott Graham, Kenneth C. Sevcik, 1984, Quantitative System Performance. Computer System Analysis Using Queueing Network Models, Prentice-Hall, Inc.
2. S. Athuraliya , VH. Li, SH. Low, Q. Yin., 2001. REM: Active Queue Management, IEEE Network, 15(3), pp 48-53.
3. S. Floyd , G. Ramakrishna, S. Shenker, 2001. Adaptive RED: an Algorithm for Increasing the Robustness of RED's Active Queue Management. Technical Report, ICSI.
4. J. Aweya, M. Ouelllette, DY. Moutuno, 2001. A Control Theoretic Approach to Active Queue Management. Computer Netwowk, 36 (2-3), pp 203-235.
5. AJ Hussein, W. Mike, T. Fadi, AA Amer, 2007. Performance Evaluation for DRED Discret Time Queueing Network Analytical Model. Journal of Netwowk and Computer Application, 31, pp 750-770.
6. Lê Quyết Thắng và Phạm Nguyên Khang, 2013. Lý thuyết xếp hàng và ứng dụng đánh giá hệ thống, Nhà xuất bản Đại học Cần thơ.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

TIẾP CẬN LUỒNG CỰC ĐẠI TRONG MẠNG CHO BÀI TOÁN XẾP LỊCH BIỂU

Phạm Nguyên Khang[1], Võ Trí Thức[1], Nguyễn Bá Diệp[1] và Bùi Lê Diễm[2]

[1] *Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ*
[2] *Khoa Sư phạm, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Maximum flow network approach for scheduling problem

Từ khóa:
Xếp lịch biểu, luồng cực đại, chi phí cực tiểu, tối ưu hoá

Keywords:
Scheduling problem, maximum flow, minimum cost, optimization

ABSTRACT

In this paper, we present a solution for the scheduling problem of pratical courses in Can Tho University using Maximum Flow Network approach. The scheduling problem for practical courses is concerned to assign students to groups/practical rooms under some constraints such as room capacity, teachers' schedule, etc. The problem attempts to optimize the performance criteria and distribute the students fairly to rooms depending on the ratio of room capacity and the number of students enrolled. We model the scheduling problem as a maximum flow problem where each student is modeled as a source node and each room as a sink node. Capacity constraints are represented as maximum capacity of edges connecting source nodes and sink nodes. We have also added constraints on minimum number of students assigned to a room as minimum capacity of edges. The approach is implemented on Google computing platform using Google Apps Script. A case study is experimented by enrollment data of practical courses in Can Tho University. Results show that our solution is tractable.

TÓM TẮT

Trong bài viết này, chúng tôi trình bày một giải pháp cho bài toán xếp lịch các học phần thực hành tại các trường đại học sử dụng mô hình luồng cực đại trong mạng. Bài toán xếp lịch thực hành là một dạng của bài toán xếp thời khoá biểu tổng quát trong đó liên quan đến việc phân các sinh viên vào các nhóm/phòng thực hành sao cho thoả mãn các ràng buộc về lịch rảnh của sinh viên, giảng viên, sức chứa của phòng và quan trọng nhất là khai thác tối đa hiệu suất sử dụng của các phòng thực hành. Với các ràng buộc như trên, bài toán xếp lịch thực hành có thể được mô hình hoá về bài toán tìm luồng cực đại trong mạng với độ phức tạp giải thuật đa thức. Mỗi sinh viên là một đỉnh phát, mỗi phòng thực hành là một đỉnh thu. Sức chứa của phòng có thể được ràng buộc bằng khả năng thông qua của cung tương ứng. Ta cũng có thể thêm vào ràng buộc số lượng sinh viên tối thiểu cho một phòng bằng các cách đặt cận dưới cho các cung. Vấn đề tối ưu tỉ lệ giữa sức chứa của phòng và số sinh viên được gán vào phòng được giải quyết bằng tiếp cận bài toán luồng cực đại với chi phí thấp nhất (minimum cost maximum flow). Giải pháp được triển khai bằng công nghệ điện toán đám mây của Google sử dụng ngôn ngữ Google Apps Script. Kết quả thực nghiệm trên dữ liệu Đăng ký thực hành cho học phần Thực hành Tin học căn bản của bộ môn Sư phạm Toán, Khoa Sư phạm cho thấy rằng giải pháp mà chúng tôi đề xuất là phù hợp.

1 GIỚI THIỆU

Bài toán xếp thời khóa biểu là một trường hợp cụ thể của bài toán lập kế hoạch được ứng dụng rộng rãi trong nhiều lĩnh vực. Đây là bài toán được xếp vào lớp NP-khó (không tồn tại giải thuật tìm được lời giải tối ưu trong thời gian đa thức). Để tìm được lời giải tối ưu thường mất rất nhiều thời gian nhất là khi gia tăng số lượng môn học, giáo viên và phòng. Hiện nay có hai giải pháp chính được sử dụng để giải bài toán này: sử dụng tiếp cận tìm kiếm cục bộ (local search) [2, 9, 10] và giải pháp xếp lịch bán tự động. Tuy nhiên do bản chất của giải thuật, các tiếp cận tìm kiếm cục bộ không thể cho lời giải tối ưu. Wasfy và các cộng sự đã trình bày trong [13] một tiếp cận để giải quyết bài toán xếp lịch biểu bằng phương pháp quy hoạch nguyên (ILP) [1] sử dụng bộ phương pháp thoả mãn ràng buộc (SAT) [6, 8, 11, 12]. Mặc dù phương pháp quy hoạch nguyên sử dụng thoả mãn ràng buộc có nhiều thành công trong việc giải một số bài toán, nhưng các giải thuật ILP vẫn sử dụng heuristics để tìm kiếm lời giải nên kết quả có thể sẽ không tối ưu. Giải pháp bán tự động thường phải kết hợp với một vài phương pháp hiển thị để giúp người sử dụng tự mình xếp lịch biểu tốt. Tuy nhiên nhược điểm chính của phương pháp này là người sử dụng phải thực hiện rất nhiều thao tác để có được một lịch biểu hoàn chỉnh (chương trình máy tính chỉ hỗ trợ kiểm tra trùng lịch).

Từ khi Trường Đại học Cần Thơ quyết định thay đổi toàn diện và triệt để sang đào tạo theo hệ thống tín chỉ, việc xếp thời khóa biểu các học phần đã được phân cấp về các Khoa. Giáo vụ khoa chịu trách nhiệm xếp thời khóa biểu cho các học phần mà khoa mình phụ trách. Xét về bản chất việc làm này đã giảm kích thước đầu vào của bài toán xếp lịch biểu. Với sự trợ giúp của hệ thống xếp lịch biểu đang được sử dụng, việc xếp thời khóa biểu cho các học phần lý thuyết được giải quyết khá thuận lợi, dễ dàng. Tuy nhiên, đối với các học phần thực hành, hoặc các học phần vừa có lý thuyết vừa có thực hành đều không được xếp lịch thực hành trước. Lý do cơ bản là lịch thực hành sẽ phụ thuộc rất nhiều số lượng sinh viên đăng ký môn học. Nếu xếp lịch trước có thể dẫn đến việc sử dụng không hiệu quả các phòng thực hành (do số sinh viên đăng ký quá ít chẳng hạn). Do đó, hiện nay việc xếp lịch thực hành thường chỉ được thực hiện sau khi sinh viên đã đăng ký các học phần và có thời khóa biểu của các học phần lý thuyết. Đối với các Khoa có số học phần có thực hành nhiều như: Khoa Sư phạm, Khoa Khoa học Tự nhiên và Khoa Công nghệ Thông tin và Truyền thông (thực hành tin học), Khoa Công nghệ (thực hành kỹ thuật điện), việc xếp lịch thực hành cho vài trăm sinh viên trở thành vấn đề. Các giải pháp hiện nay được các Khoa thực hiện có thể phân thành 2 nhóm. Nhóm thứ nhất, thực hiện thủ công: giáo viên phân nhóm và dán thông báo, sinh viên đăng ký trực tiếp trên giấy và tùy vào tình hình đăng ký mà giáo viên phụ trách giảng dạy sẽ xếp lịch. Nhóm thứ hai cho sinh viên đăng ký thông qua một hệ thống nào đó (ví dụ một website tự tạo). Sau khi đã có nhu cầu đăng ký, việc xếp lịch được thực hiện thủ công. Tóm lại, cho dù cho sinh viên đăng ký bằng cách nào đi nữa thì việc xếp lịch sau cùng vẫn phải thực hiện thủ công dẫn đến mất nhiều thời gian công sức thậm chí có thể tồn tại sai sót trong khi xếp lịch.

Trong bài báo này, chúng tôi đề xuất một giải pháp xếp lịch thực hành cho các học phần có thực hành tại Trường Đại học Cần Thơ. Giải pháp đề xuất sẽ thực hiện việc phân tự động sinh viên vào các nhóm thực hành phù hợp (không bị trùng lịch) và sử dụng các phòng thực hành một cách hiệu quả (khai thác tối đa hiệu suất sử dụng của phòng). Các đóng góp chính của bài báo gồm: (i) đề xuất quy trình đăng ký và xếp lịch biểu, (ii) mô hình hoá bài toán xếp lịch biểu về bài toán tìm luồng cực đại trên mạng, (iii) đề xuất hàm mục tiêu để tối ưu hiệu quả sử dụng các phòng thực hành cùng với giải thuật cải tiến cho bài toán luồng cực đại với chi phí thấp nhất trong bài toán xếp lịch biểu. Ngoài ra, các lựa chọn của sinh viên cũng sẽ được xem xét ưu tiên theo thứ tự. Ví dụ như: lựa chọn 1 sẽ được xem xét trước lựa chọn 2.

Phần tiếp theo của bài viết được trình bày như sau: phần 2 trình bày ngắn gọn về Bài toán xếp lịch biểu; phần 3 trình bày mạng và luồng cực đại trong mạng; phần 4 trình bày ứng dụng của luồng cực đại trong mạng cho bài toán xếp lịch biểu. Kết quả thực nghiệm được trình bày trong phần 5, tiếp theo là kết luận và hướng phát triển.

2 GIẢI PHÁP CHO BÀI TOÁN XẾP LỊCH BIỂU

Bài toán xếp lịch biểu chúng tôi đề cập đến trong bài viết này phát sinh từ nhu cầu thực tế ở Khoa CNTT&TT, Bộ môn Toán Khoa Sư phạm và Khoa Công nghệ, trường Đại học Cần Thơ. Các học phần thực hành (hoặc học phần vừa có lý thuyết vừa có thực hành) không được xếp thời biểu trước mà giảng viên (hoặc giảng giáo viên) phụ trách học phần phải tự xếp lịch. Quy trình xếp lịch thực hành cho mỗi học phần như sau: (i) với số lượng sinh viên đăng ký tham gia học phần đã biết, giảng viên ước lượng số lượng nhóm sao cho phù

hợp số lượng sinh viên và sức chứa của các phòng thực hành (phòng máy tính, phòng thí nghiệm); (ii) chọn buổi thực hành cho các nhóm và phân công giảng viên hướng dẫn thực hành (iii) thông báo cho sinh viên đăng ký nhóm và (iv) giảng viên phân sinh viên vào các nhóm. Giai đoạn (i) được thực hiện khá dễ dàng dựa vào số lượng sinh viên và sức chứa các phòng thực hành. Việc chọn buổi thực hành và phân công giảng viên được thực hiện dựa trên lịch rảnh của giáo viên và của phòng thực hành. Giai đoạn cho sinh viên đăng ký nhóm là công việc khá phức tạp. Giai đoạn cuối cùng là giai đoạn khó nhất: phân sinh viên vào các nhóm sao cho số lượng sinh viên của mỗi nhóm không quá lớn (vượt quá khả năng của phòng thực hành tương ứng với nhóm) và cũng không quá nhỏ (khai thác tối đa hiệu suất sử dụng phòng thực hành). Với quy trình xếp lịch như thế, chúng ta đã chia nhỏ bài toán xếp thời khóa biểu tổng quát (rất khó) thành các bài toán nhỏ (dễ hơn).

3 MẠNG VÀ LUỒNG CỰC ĐẠI TRONG MẠNG

Mạng (network) là một đồ thị có hướng G = <V, E> với tập đỉnh V và tập cung E. Trên tập cung E ta định nghĩa hàm *khả năng thông qua* (capacity function) có giá trị không âm c: E → $\mathbf{R}^+$. Để đơn giản, ta có thể giả sử G không chứa đa cung. Nếu có một cung đi từ đỉnh u đến đỉnh v, cung này sẽ là cung duy nhất và được ký hiệu là (u, v). Trên mạng ta phân biệt 2 đỉnh: *đỉnh phát* (source) s và *đỉnh thu* (sink) t.

Luồng (flow) là một hàm f: E → $\mathbf{R}^+$ thỏa mãn các ràng buộc sau:

– Luồng trên cung không thể vượt quá khả năng thông qua của cung:

$$f(u,v) \le c(u, v) \quad (1)$$

– Bảo toàn luồng: tổng luồng đi vào một đỉnh bằng tổng luồng đi ra khỏi đỉnh đó: $\forall u \notin \{s,t\}$,

$$\sum_{(k,u)\in E} f(k,u) = \sum_{(u,v)\in E} f(u,v) \quad (2)$$

Ngoài ra, ta còn có:

– Tổng luồng đi ra đỉnh phát bằng tổng luồng đi vào đỉnh thu

$$\sum_{(s,u)\in E} f(s,u) = \sum_{(v,t)\in E} f(v,t) \quad (3)$$

– Giá trị luồng (value of flow), kí hiệu $|f|$ được định nghĩa bằng tổng luồng đi ra khỏi đỉnh phát hoặc tổng luồng đi vào đỉnh thu:

$$|f| = \sum_{(s,u)\in E} f(s,u) = \sum_{(v,t)\in E} f(v,t) \quad (4)$$

Bài toán tìm luồng cực đại (trên mạng) là bài toán cực đại hóa $|f|$ với các ràng buộc trên. Giải thuật đầu tiên được sử dụng để tìm luồng cực đại là giải thuật *đường tăng luồng* (augmenting path algorithm) của Ford & Fullkerson [5, 8]. Kế đến, Dinic [3] và Edmonds & Karp [4] đã độc lập cải tiến giải thuật đường tăng luồng để có được giải thuật có độ phức tạp thời gian đa thức. Kể từ đó, đã có rất nhiều giải thuật hiệu quả hơn được phát triển có thể kể đến như phương pháp *đẩy/gán nhãn lại* (push/relabel) của Golberg & Tarjan [7].

4 ỨNG DỤNG LUỒNG CỰC ĐẠI CHO BÀI TOÁN XẾP LỊCH BIỂU

Như đã phân tích ở trên, bài toán xếp lịch thực hành được quy về các bài toán con đơn giản hơn. Bài toán phân nhóm cho sinh viên là bài toán khó nhất và đáng được quan tâm. Trong phần này chúng tôi sẽ mô hình hóa bài toán phân nhóm về bài toán tìm luồng cực đại trên mạng.

Gọi:

– S = $\{u_1, u_2, \ldots, u_n\}$ là tập các sinh viên

– G = $\{v_1, v_2, \ldots, v_m\}$ là tập các nhóm

– s: đỉnh phát

– t: đỉnh thu

– $h(v)$, $v \in$ G: sĩ số tối đa của nhóm (ứng với sức chứa của phòng)

– $l(v)$, $v \in$ G: sĩ số tối thiểu của nhóm (ứng với sức chứa của phòng)

$d(u,v)$, $u \in$ S, $v \in$ G: hàm đăng ký nhóm của sinh viên. $d(u,v) = 1$ nếu sinh viên u đăng ký nhóm v, ngược lại $d(u,v) = 0$.

Đồ thị G = <V,E> của mạng sẽ là:

– V = S ∪ G ∪ $\{s, t\}$

– E = $\{(s,u) \mid u \in S\} \cup \{(u, v) \mid u \in S, v \in G, d(u,v) = 1\} \cup \{(v, t) \mid v \in G\}$

– $\forall u \in$ S: $c(s, u) = 1$

– $\forall (u, v) \in$ E, $u \in$S, v∈G: $c(u,v) = 1$

– $\forall v \in$ G, $c(v, t) = h(v)$

Mỗi sinh viên được mô hình thành một *đỉnh-sinh viên*, mỗi nhóm cũng tương ứng với một *đỉnh-nhóm*. Ta thêm vào một đỉnh phát s và một đỉnh thu t.

Nếu sinh viên u_i đăng ký nhóm v_j sẽ có một cung nối từ *đỉnh-sinh viên* u_i đến *đỉnh-nhóm* v_j, khả

năng thông qua của cung này là 1 (mỗi sinh viên chỉ được phân vào 1 nhóm). Thêm các cung nối từ đỉnh phát *s* đến các đỉnh-sinh viên, khả năng thông qua của cung này cũng là 1 (mỗi sinh viên chỉ được đăng ký tối đa 1 nhóm). Sau cùng, thêm các cung nối các đỉnh-nhóm $v \in G$ đến *t* với khả năng thông qua là $h(v)$. Mạng kết quả được minh họa trong Hình 1.

4.1 Ràng buộc sĩ số tối thiểu

Để tránh trường hợp một nhóm có quá ít sinh viên ta có thể bổ sung thêm ràng buộc sĩ số tối thiểu của một nhóm: tổng số sinh viên của nhóm *v* không được nhỏ hơn $l(v)$. Với mô hình luồng cực đại trên mạng, ràng buộc này có thể dễ dàng tích hợp vào mô hình: với mỗi cung, ngoài khả năng thông qua lớn nhất ta bổ sung thêm khả năng thông qua nhỏ nhất. Cách đơn giản nhất để giải bài toán luồng cực đại trên mạng với ràng buộc khả năng thông qua nhỏ nhất của các cung là biến đổi về bài toán luồng cực đại chuẩn. Ta xây dựng đồ thị G' = <V', E'> từ đồ thị G = <V, E>:

– Thêm vào hai đỉnh mới *s*' và *t*'.

– Thêm cung nối *s*' đến tất cả các đỉnh trong V.

– Thêm cung nối các đỉnh trong V đến *t*'.

– Thêm cung nối *t* đến *s*.

– Khả năng thông qua mới của các cung được định nghĩa như sau:

- Với mỗi đỉnh $v \in V$, đặt:
- $c'(s', v) = \sum_{u \in V} d(u,v)$ và
- $c'(v, t') = \sum_{w \in V} d(v,w)$
- Với mỗi cung (u, v) $\in$ E, đặt c'(u,v) = $c(u,v) - d(u,v)$
- Đặt $c'(t, s) = \infty$

Một cách trực quan, đồ thị G' được xây dựng bằng cách thay mỗi cung (*u*, *v*) trong đồ thị G bằng 3 cung:

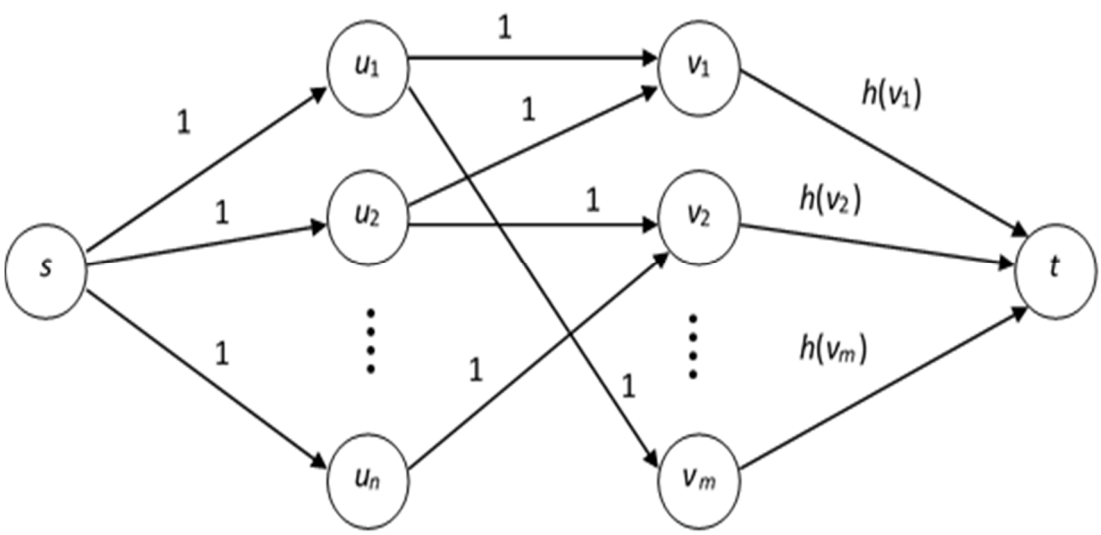
– (*u*, *v*) với khả năng thông qua $c(u, v) - d(u, v)$;

– (*s*', *v*) với khả năng thông qua $d(u, v)$ và

– (*u*, *t*') với khả năng thông qua $d(u, v)$

Trong G', tổng khả năng thông qua của các cung xuất phát từ *s*' bằng tổng khả năng thông qua của cả các cung đi vào *t*' và bằng $\sum_{(u,v) \in E} d(u,v)$. Ta gọi luồng có giá trị đúng bằng $\sum_{(u,v) \in E} d(u,v)$ là *luồng bão hòa* (saturating flow) vì nó bão hòa tất cả các cung rời khỏi *s*' hoặc đi vào *t*'. Nếu G' có luồng bão hòa, nó sẽ có luồng cực đại.

Người ta cũng đã chứng minh rằng: *Mạng G sẽ có luồng khả thi (feasible flow) khi và chỉ khi G' có luồng bão hòa.*



Hình 1: Mạng phân nhóm sinh viên

Giải thuật tìm luồng cực đại trong trường hợp khả năng thông qua của cung bị chặn dưới theo các bước như sau:

1. Xây dựng đồ thị G' và tìm luồng cực đại f' từ s' đến t' trên G'. Nếu tất cả các cung ra khỏi s' đều bão hòa (luồng đi qua cung này bằng khả năng thông qua lớn nhất của nó), ta tìm được *luồng khả thi* (admissible flow) f_a trên G:

   $$f_a(u,v) = f'(u,v) + d(u,v)$$

   với mọi cung $(u,v) \in E$

2. Trên G, ta định nghĩa lại khả năng thông qua lớn nhất của các cung như sau:

   - $c_r(u,v) = c(u,v) - f_a(u,v)$

3. Tìm luồng cực đại f_r từ s đến t trên G với khả năng thông qua lớn nhất c_r.

4. Luồng cực đại thỏa mãn ràng buộc khả năng thông qua lớn nhất và nhỏ nhất là:

   - $f(u,v) = f_a(u,v) + f_r(u,v)$

4.2 Tối ưu hoá hiệu quả sử dụng phòng

Như đã trình bày ở trên, mỗi nhóm được gán cho một phòng thực hành. Sĩ số tối đa của một nhóm thường là sức chứa của phòng thực hành tương ứng. Do đó, ngoài ràng buộc về sĩ số thấp nhất, ta cũng muốn số lượng sinh viên được xếp vào nhóm phải tương ứng với sĩ số tối đa của nhóm (ứng với sức chứa của phòng thực hành). Không xếp quá ít sinh viên vào một nhóm có sĩ số tối đa quá lớn. Bằng cách này, sẽ không có phòng quá tải trong khi các phòng khác thì quá trống. Như thế việc sử dụng phòng sẽ đạt hiệu quả hơn. Ràng buộc này có thể được mô hình hóa một cách đơn giản như sau:

$$\sum_{v \in G} \frac{c(v,t)}{f(v,t)} \rightarrow \min \quad (5)$$

Đó là tỉ số giữa sĩ số tối đa và số sinh viên thực sự được xếp vào nhóm v phải nhỏ nhất có thể. Như thế bài toán trở thành tìm luồng cực đại trên mạng thỏa mãn ràng buộc (5). Đây có thể xem như một biến thể của bài toán *luồng cực đại với giá thành cực tiểu* (minimum-cost maximum-flow). Giải thuật gốc cho bài toán *luồng cực đại với giá thành cực tiểu* được mô tả như sau:

1. *Tìm luồng cực đại **f** trong mạng **G***
2. *Tìm chu trình giảm giá thành và cập nhật lại luồng **f**.*
3. *Lặp lại bước 2 cho đến khi không thể tìm được chu trình giảm giá thành nào nữa.*

Do tính chất đặc biệt của bài toán phân nhóm sinh viên: có cấu trúc đồ thị đơn giản và hàm mục tiêu cần tối ưu khác với hàm mục tiêu của bài toán *luồng cực đại với giá thành cực tiểu*, chúng tôi đề xuất giải thuật giải quyết bài toán như sau:

1. Tìm luồng cực đại f trong mạng G
2. Với mỗi đỉnh-sinh viên $u \in S$

   Xét các cặp $v, w \in G$ có cung nối từ u đến v và cung nối từ u đến w

   Nếu $f(u,v) = 1$ và $f(u,w) = 0$

   Nếu việc đổi luồng làm giảm hàm mục tiêu

   Cập nhật lại luồng f.

   Nếu $f(u,w) = 1$ và $f(u,v) = 0$

   Nếu việc đổi luồng làm giảm hàm mục tiêu

   Cập nhật lại luồng f.

3. Lặp lại bước 2 cho đến khi không thể tìm được chu trình giảm giá thành nào nữa.

Một ví dụ về tìm chu trình giảm giá thành được minh hoạ trong Hình 2. Chu trình tìm thấy là u → w → t → v → u. Đường liền nét là tăng luồng và đường không liền nét là đường giảm luồng.

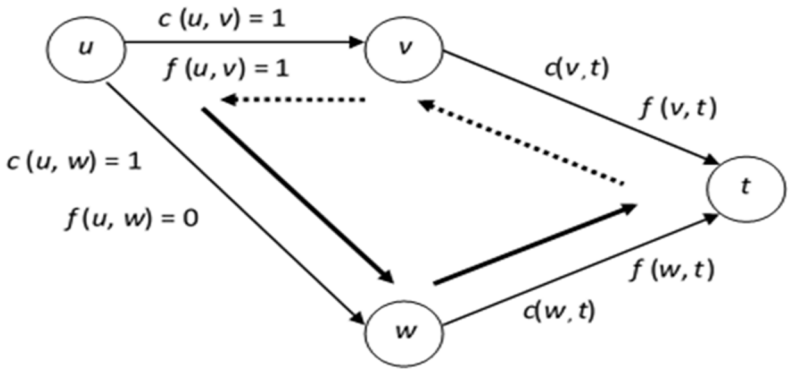
5 KẾT QUẢ THỰC NGHIỆM

Từ các phân tích được trình bày ở trên, giải pháp tổng thể cho quy trình xếp lịch được mô tả ngắn gọn như sau:

1. Chuẩn bị nhóm/phòng
2. Cho sinh viên đăng ký
3. Thu thập dữ liệu đăng ký và tiền xử lý
4. Xây dựng mạng
5. Tìm luồng cực đại trên mạng
6. Gán sinh viên vào nhóm dựa trên kết quả của luồng cực đại.

Chúng tôi đã triển khai quy trình đăng ký và xếp lịch thực hành trên trên nền tảng công nghệ điện toán đám mây của Google với ngôn ngữ lập trình ngữ Google Apps Script. Tất cả các thứ cần chuẩn bị là tạo một bảng tính (spreadsheet) dùng

làm nơi lưu trữ dữ liệu đăng ký, các thông số của hệ thống (số nhóm, sức chứa), các tham số của các giải thuật và cũng là nơi lưu kết quả xếp lịch. Việc tổ chức cho sinh viên đăng ký được thực hiện khá đơn giản với Google Form (bước 2). Tạo một biểu mẫu đăng ký với Google Form và gởi địa chỉ của biểu mẫu cho sinh viên. Một ví dụ về biểu mẫu đăng ký được trình bày trong Hình 3. Trong biểu mẫu này, phần quan trọng nhất là mã số sinh viên và các lựa chọn của sinh viên. Tuỳ vào lịch rảnh của mình sinh viên sẽ lựa chọn các nhóm có lịch phù hợp. Để bài toán dễ có lời giải, có thể khuyến khích sinh viên lựa chọn nhiều nhóm khác nhau. Trong thời hạn đăng ký, sinh viên có thể điều chỉnh các lựa chọn. Sau khi sinh viên đăng ký xong, dữ liệu đăng ký được tự động lưu trong bảng tính (bước 3).

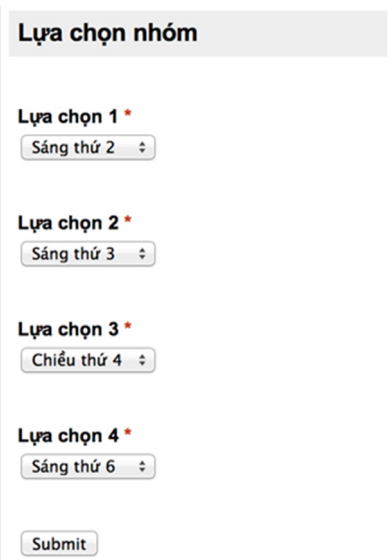


Hình 2: Tìm chu trình giảm giá thành trên mạng phân nhóm sinh viên

Một phần của dữ liệu đăng ký thực hành môn Tin học căn bản ở Khoa Sư phạm, Đại học Cần Thơ được minh hoạ trong Bảng 1. Dữ liệu sẽ được xử lý trước khi chuyển qua bước xây dựng mạng và tìm luồng cực đại trong mạng. Trong phần tiền xử lý này, ta loại bỏ các sinh viên không có danh sách. Nếu sinh viên đăng ký nhiều lần ta ghi nhận lần đăng ký sau cùng và bỏ qua các lần đăng ký trước đó. Ngoại trừ các bước 1 và 2 ta phải thực hiện thủ công, các bước còn lại được lập trình để có thể thực hiện tự động. Bước 2 cũng có thể được thực hiện tự động nếu hệ thống được cung cấp thông tin về lịch rảnh của sinh viên.

Sau khi dữ liệu đã được làm "sạch", ta thiết lập các tham số cho hệ thống và giải thuật để bắt đầu xây dựng mạng và tìm luồng cực đại. Đầu tiên, có thể cho sức chứa tối thiểu của tất cả các nhóm bằng 0 và thực hiện việc xếp lịch. Tuỳ vào kết quả xếp nhóm, ta có thể điều chỉnh sức chứa tối thiểu của các nhóm. Ngoài ra ta cũng có thể bật/tắt chức năng tối ưu tỉ lệ của sức chứa tối đa và số sinh viên gán vào các nhóm. Nếu chức năng này được bật, hệ thống sẽ tìm luồng cực đại đồng thời tối ưu hàm mục tiêu (5).

Bảng 2 trình bày kết quả thu được khi kiểm thử hệ thống xếp lịch cho học phần thực hành tin học căn bản với số sinh viên đăng ký 694 trong đó 240 trường hợp ký trùng hoặc không có tên trong danh sách. Sau khi tiền xử lý (tự động) số sinh viên còn lại là 480 được xếp vào 12 nhóm. Trường hợp không tối ưu tỉ lệ giữa sức chứa tối đa và số sinh viên được gán vào nhóm, sẽ có trường hợp 1 số nhóm có rất nhiều sinh viên, trong khi các nhóm khác có rất ít sinh viên (so với sức chứa tối đa của nhóm). Hàm mục tiêu (5) có giá trị 18.16. Trong trường hợp có tối ưu tỉ lệ sức chứa, ta thu được giá trị 16.25. Kết quả xếp lịch được lưu trở lại bảng tính (xem Bảng 3) và có thể sử dụng để in kết quả theo từng nhóm như Bảng 4.



Hình 3: Một phần của Biểu mẫu đăng ký nhóm

Bảng 1: Dữ liệu đăng ký thực hành học phần tin học căn bản

Timestamp	MSSV	Họ và tên	Lựa chọn 1	Lựa chọn 2	Lựa chọn 3	Lựa chọn 4
02/01/2013 14:57:05	B1200392	Nguyễn Dương Thanh Phú	Chiều thứ 2	Sáng thứ 3	Chiều thứ 5	Sáng thứ 5
02/01/2013 15:19:57	B1200352	Trương Gia Đại	Sáng thứ 6	Chiều thứ 6	Chiều thứ 7	Sáng thứ 7
02/01/2013 15:22:12	B1200962	Trương Thị Kiều Trang	Chiều thứ 2	Sáng thứ 6	Sáng thứ 2	Chiều thứ 6
02/01/2013 15:23:21	B1200467	Nguyễn Phi Vân	Sáng thứ 3	Chiều thứ 3	Sáng thứ 7	Chiều thứ 6
02/01/2013 15:27:56	B1200453	Lâm Minh Tài	Sáng thứ 3	Chiều thứ 6	Chiều thứ 7	Sáng thứ 7
02/01/2013 15:29:08	B1200370	Nguyễn Tùng Lâm	Chiều thứ 2	Sáng thứ 3	Chiều thứ 5	Sáng thứ 4
02/01/2013 15:33:01	B1207990	Dương Quang Khang	Sáng thứ 3	Chiều thứ 5	Sáng thứ 7	Chiều thứ 6
02/01/2013 15:34:09	B1208023	Phạm Thị Cẩm Tiên	Sáng thứ 3	Chiều thứ 6	Chiều thứ 7	Sáng thứ 7
02/01/2013 15:35:05	B1200471	Nguyễn Nguyên Bá	Sáng thứ 3	Sáng thứ 3	Sáng thứ 7	Chiều thứ 6
02/01/2013 15:52:12	1117555	Hồ Trọng Nhân	Chiều thứ 4	Chiều thứ 3	Chiều thứ 7	Chiều thứ 6
02/01/2013 16:52:16	B1208188	Võ Văn Vũ	Sáng thứ 5	Chiều thứ 5	Sáng thứ 2	Sáng thứ 7
02/01/2013 16:55:49	B1200949	Châu Thị Kim Thi	Sáng thứ 5	Sáng thứ 4	Chiều thứ 5	Chiều thứ 4
02/01/2013 17:05:35	B1200973	Văn Thị Như Yến	Chiều thứ 2	Sáng thứ 6	Sáng thứ 6	Chiều thứ 2
02/01/2013 17:07:38	6116284	Nguyễn Hữu Thịnh	Sáng thứ 6	Sáng thứ 6	Sáng thứ 6	Sáng thứ 6
02/01/2013 17:09:33	6116251	Lý Tuyết Kha	Sáng thứ 6	Sáng thứ 6	Sáng thứ 6	Sáng thứ 6
02/01/2013 17:13:07	B1200386	Phạm Thị Hồng Nhãn	Chiều thứ 2	Sáng thứ 3	Chiều thứ 5	Sáng thứ 5
02/01/2013 17:24:25	B1200421	LÊ MỘNG TUYỀN	Chiều thứ 2	Sáng thứ 3	Sáng thứ 2	Sáng thứ 5
02/01/2013 17:29:45	B1200386	Phạm Thị Hồng Nhãn	Chiều thứ 2	Sáng thứ 3	Chiều thứ 5	Sáng thứ 5
02/01/2013 17:33:29	b1200394	QUÁCH THU PHƯƠNG	Chiều thứ 2	Sáng thứ 3	Sáng thứ 7	Sáng thứ 5
02/01/2013 17:34:29	B1200421	LÊ MỘNG TUYỀN	Chiều thứ 2	Sáng thứ 3	Sáng thứ 7	Sáng thứ 5

Bảng 2: Kết quả xếp lịch trong trường hợp không tối ưu và tối ưu tỉ lệ sức chứa tối đa và số sinh viên trong nhóm

STT	Nhóm	Sức chứa tối đa	Sức chứa tối thiểu	Không tối ưu		Tối ưu theo hàm mục tiêu (5)	
				SV/Nhóm	Tỉ lệ	SV/Nhóm	Tỉ lệ
1	Sáng thứ 2	75	40	40	1.88	40	1.88
2	Chiều thứ 2	75	0	74	1.01	45	1.67
3	Sáng thứ 3	40	39	39	1.03	39	1.03
4	Chiều thứ 3	72	0	35	2.06	44	1.64
5	Sáng thứ 4	40	0	25	1.60	33	1.21
6	Chiều thứ 4	40	0	28	1.43	33	1.21
7	Sáng thứ 5	75	70	70	1.07	70	1.07
8	Chiều thứ 5	75	0	55	1.36	45	1.67
9	Sáng thứ 6	40	0	40	1.00	32	1.25
10	Chiều thứ 6	40	0	40	1.00	33	1.21
11	Sáng thứ 7	40	0	18	2.22	33	1.21
12	Chiều thứ 7	40	0	16	2.50	33	1.21
Tổng					**18.16**		**16.25**

Bảng 3: Kết quả việc phân sinh viên vào các nhóm

Timestamp	MSSV	Họ và tên	Email	Kết quả xếp nhóm	Lựa chọn 1	Lựa chọn 2	Lựa chọn 3	Lựa chọn 4
02/01/2013 15:22:12	B1200962	Trương Thị Kiều Trang	01693055217	Chiều thứ 2	Chiều thứ 2	Sáng thứ 6	Sáng thứ 2	Chiều thứ 6
02/01/2013 15:23:21	B1200467	Nguyễn Phi Vân	01259595990	Sáng thứ 7	Sáng thứ 3	Chiều thứ 3	Sáng thứ 7	Chiều thứ 6
02/01/2013 15:27:56	B1200453	Lâm Minh Tài	0985886645	Sáng thứ 7	Sáng thứ 3	Chiều thứ 6	Chiều thứ 7	Sáng thứ 7
02/01/2013 15:29:08	B1200370	Nguyễn Tùng Lâm	0939707015	Sáng thứ 4	Chiều thứ 2	Sáng thứ 3	Chiều thứ 5	Sáng thứ 4
02/01/2013 15:33:01	B1207990	Dương Quang Khang	0919493494	Chiều thứ 6	Sáng thứ 3	Chiều thứ 5	Sáng thứ 7	Chiều thứ 6
02/01/2013 15:34:09	B1208023	Phạm Thị Cẩm Tiên	0917525322	Sáng thứ 7	Sáng thứ 3	Chiều thứ 6	Chiều thứ 7	Sáng thứ 7
02/01/2013 15:35:05	B1200471	Nguyễn Nguyên Bá	01655365883	Sáng thứ 7	Sáng thứ 3	Sáng thứ 3	Sáng thứ 7	Chiều thứ 6
02/01/2013 15:52:12	1117555	Hồ Trọng Nhân	nhan117555@stude	BỎ QUA	Chiều thứ 4	Chiều thứ 3	Chiều thứ 7	Chiều thứ 6
02/01/2013 16:52:16	B1208188	Võ Văn Vũ	01286521844	BỎ QUA	Sáng thứ 5	Chiều thứ 5	Sáng thứ 2	Sáng thứ 7
02/01/2013 16:55:49	B1200949	Châu Thị Kim Thi	01882122021	BỎ QUA	Sáng thứ 5	Sáng thứ 4	Chiều thứ 5	Chiều thứ 4
02/01/2013 17:05:35	B1200973	Văn Thị Như Yến	0939698220	Chiều thứ 2	Chiều thứ 2	Sáng thứ 6	Sáng thứ 6	Chiều thứ 2
02/01/2013 17:07:38	6116284	Nguyễn Hữu Thịnh	01684823419	BỎ QUA	Sáng thứ 6	Sáng thứ 6	Sáng thứ 6	Sáng thứ 6
02/01/2013 17:09:33	6116251	Lý Tuyết Kha	01652238176	BỎ QUA	Sáng thứ 6	Sáng thứ 6	Sáng thứ 6	Sáng thứ 6
02/01/2013 17:13:07	B1200386	Phạm Thị Hồng Nhãn	nhanB1200386@ctu	BỎ QUA	Chiều thứ 2	Sáng thứ 3	Chiều thứ 5	Sáng thứ 5
02/01/2013 17:24:25	B1200421	LÊ MỘNG TUYỀN	01258383095	BỎ QUA	Chiều thứ 2	Sáng thứ 3	Sáng thứ 2	Sáng thứ 5
02/01/2013 17:29:45	B1200386	Phạm Thị Hồng Nhãn	nhanB1200386@stu	Chiều thứ 5	Chiều thứ 2	Sáng thứ 3	Chiều thứ 5	Sáng thứ 5
02/01/2013 17:33:29	b1200394	QUÁCH THU PHƯƠNG	01699944396	Sáng thứ 7	Chiều thứ 2	Sáng thứ 3	Sáng thứ 7	Sáng thứ 5
02/01/2013 17:34:29	B1200421	LÊ MỘNG TUYỀN	01258383095	Sáng thứ 7	Chiều thứ 2	Sáng thứ 3	Sáng thứ 7	Sáng thứ 5

Xét về mặt thời gian thực hiện giải thuật tìm luồng cực đại, với số sinh viên khoảng 600-1000, thời gian thực hiện khoảng 20-30 giây. Lý do chính là do các giải thuật được cài đặt bằng Google scripts, nên thời gian thực hiện không thể so sánh với phiên bản cài đặt trên Desktop bằng các ngôn ngữ mạnh như C/C++. Tuy nhiên, thời gian thực hiện như trên vẫn có thể chấp nhận được. Hơn nữa, việc sử dụng nền tảng của Google để triển khai giải pháp lại giúp ta tiết kiệm được thời gian và công sức để xây dựng phần thu thập số liệu đăng ký.

6 KẾT LUẬN

Chúng tôi vừa đề xuất một giải pháp xếp lịch biểu cho các môn học sử dụng tiếp cận Luồng cực đại trong mạng. Chúng tôi đã mô hình hoá bài toán xếp lịch biểu về bài toán tìm luồng cực đại trên mạng. Ràng buộc mỗi sinh viên chỉ được gán vào một nhóm được mô hình bằng khả năng thông qua của cung xuất phát từ đỉnh-sinh viên bằng 1. Trong khi đó, ràng buộc về sức chứa của nhóm được mô hình hoá về khả năng thông qua của cung (cả cận trên và cận dưới). Chúng tôi cũng đã đề xuất một hàm mục tiêu phù hợp với việc tối ưu tỉ lệ giữa sức chứa tối đa và số sinh viên vào nhóm cùng với một giải thuật tìm luồng cực đại với hàm mục tiêu nhỏ nhất. Tất cả các đề xuất đã được cài đặt trên nền tảng điện toán đám mây của Google. Chúng tôi đã thực nghiệm giải pháp đề xuất trên dữ liệu đăng ký thực hành của các sinh viên ở các Khoa của Trường Đại học Cần Thơ: Khoa Sư phạm (Tin học căn bản), Khoa Công nghệ Thông tin và Truyền thông (Xử lý ảnh), Khoa Công nghệ (Kỹ thuật điện). Kết quả cho thấy rằng giải pháp rất hiệu quả vì tính đơn giản và chi phí thấp của nó.

Dựa vào các phân tích trong phần thực nghiệm, có thể thấy rằng có thể phát triển giải pháp được đề xuất bằng cách giải quyết một số vấn đề sau: (i) Tự động hoá quá trình thu thập dữ liệu đăng ký từ thời khoá biểu các môn học lý thuyết của sinh viên, (ii) Tổng quát hoá hàm mục tiêu tối ưu sao cho có thể mô hình được nhiều điều kiện tối ưu khác nhau ví dụ như: phân nhóm sao cho số nhóm sử dụng là ít nhất, (iii) và quan tâm đến vấn đề bảo mật khi đăng ký ví dụ như sử dụng tài khoản và mật khẩu để đăng nhập mới có thể đăng ký được. Chúng tôi dự định giải quyết các vấn đề trên trong một tương lai gần.

Bảng 4: In danh sách sinh viên theo nhóm

A	B	C	D
		Nhóm: Sáng thứ 2, Phòng: , Buổi: Sáng thứ 2	
STT	**MSSV**	**Họ và tên**	**Số điện thoại/email**
1	1117535	võ thị mỹ xuyên	1643262750
2	7108781	Võ Thị Kim Ngọc	939564088
3	b1200299	Huỳnh Lâm Nhã	1654361913
4	B1200337	Nguyễn Hoàng Uyên	uyenB1200337@student.ctu.edu.vn
5	B1200373	NGUYỄN TÀI LINH	1222885531
6	b1200444	huỳnh viễm my	myb1200444@student.ctu.edu.vn
7	B1200486	Trần Đắc Mạnh	manhB1200486@ctu.edu.vn
8	b1200506	Lý Mỹ Trinh	trinhb1200506@student.ctu.edu.vn
9	b1200538	Lê Thị Son	sonb1200538@student.ctu.edu.vn
10	b1200539	Bùi Thị Thanh Tâm	tamb1200539@student.ctu.edu.vn
11	b1200543	Nguyễn Thị Phương Thùy	thuyb1200543@student.ctu.edu.vn
12	B1200563	Lê Minh Bằng	bangB1200563@student.ctu.edu.vn
13	B1200600	Mai Phước Lộc	963880394
14	b1200667	trần chúc linh	linhb1200667@student.ctu.edu.vn
15	b1200708	Phan Ngọc Hà	963880061
16	B1200712	NGUYỄN THỊ HUYỀN	1629685510

TÀI LIỆU THAM KHẢO

1. Aloul, F., A. Ramani, I. Markov, and K. Sakallah, Generic ILP Versus Specialized 0-1 ILP: An Update, in *Proc. of the Int'l Conference on Computer Aided Design*, 450-457, 2002.
2. Cosla, D., A Tabu Search Algorithm for Computing an Operational Timetable, in

European Journal of Operational Research, vol. 76, 98-110, 1994.

3. Dinic, E., A. Algorithm for solution of a problem of maximum flow in a network with power estimation. *Soviet Math. Doklady* 11: 1277–1280, 1970.
4. Edmonds, J., Karp, Richard M., Theoretical improvements in algorithmic efficiency for network flow problems. Journal of the ACM (Association for Computing Machinery) 19 (2): 248–264, 1972.
5. Ford, L. R., Fulkerson, D. R. Maximal flow through a network. Canadian Journal of Mathematics 8: 399–404, 1956.
6. E. Goldberg and Y. Novikov, "BerkMin: A Fast and Robust SAT-solver," in *Proc. of the Design Automation and Test Conference in Europe* (DATE), 142-149, 2002.
7. Goldberg, Andrew V., Robert E. Tarjan. A new approach to the maximum flow problem. *Annual ACM Symposium on Theory of Computing* , 136–146. ISBN 0-89791-193-8, 1986.
8. Heineman, George T., Gary Pollice, and Stanley Selkow. Chapter 8:Network Flow Algorithms. *Algorithms in a Nutshell*. Oreilly Media. p. 226–250, 2008.
9. Hertz, A., Tabu Search for Large Scale Timetabling Problems, in *European Journal of Operation Research*, 54(1), 39-47, 1991.
10. Mooney, E., R. Dargen, and W. Parameter, Large-Scale Classroom Scheduling, in *IIE Trans.*, 28(5), 369-378, 1996.
11. Moskewicz, K., C. Madigan, Y. Zho, and S. Malik, Chaff: Engineering an efficient SAT solver, in *Proc. of the Design Automation Conference* (DAC)*,* 503-535, 2001.
12. Tam V., and D. Ting, Combining the Min-Conflicts and Look- Forward Heuristics to Effectively Solve a Set of Hard University Timetabling Problems, in *Proc. of the IEEE International Conference on Tools with Artificial Intelligence,* 2003.
13. Wasfy A. and Fadi A. Aloul, Solving the University Class Scheduling Problem Using Advanced ILP Techniques, in *Proc. of IEEE GCC'07*, 2007.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

KHAI KHOÁNG VÀ ỨNG DỤNG CỦA MẪU EPISODE MỞ RỘNG TRÊN DỮ LIỆU PHỤ THUỘC THỜI GIAN

Nguyễn Bá Diệp[1], Phạm Nguyên Khang[1], Trần Nguyễn Minh Thư[1] và Huỳnh Xuân Hiệp[2]

[1] *Bộ môn Khoa học Máy Tính, Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ*
[2] *Bộ môn Công nghệ Phần mềm, Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Extended episode mining and its application in time-related data

Từ khóa:
Dữ liệu phụ thuộc thời gian, khai khoáng mẫu episode, mẫu episode mở rộng

Keywords:
Data mining in time-related data, episode mining, extended episode

ABSTRACT

In this paper, we introduce extended episode model upgraded from episode pattern in time-related data. Based on this model, we present an algorithm that finds all frequently extended episodes from an input event sequence without rescaning. By application using new characteristics of mined extended episodes, we propose an application in the diabetes data. Experimental results of this article show that the extended episodes contain useful information for prediction models.*

TÓM TẮT

*Trong bài viết này, chúng tôi giới thiệu mô hình mẫu episode mở rộng phát triển từ mô hình mẫu episode với dữ liệu phụ thuộc vào thời gian. Dựa trên mô hình vừa trình bày, chúng tôi giới thiệu giải thuật khai khoáng mẫu episode mở rộng chỉ duyệt qua chuỗi dữ liệu sự kiện 1 lần. Sử dụng những đặc tính đặc trưng của mẫu episode mở rộng chúng tôi đề xuất một ứng dụng dự đoán trên tập dữ liệu tiểu đường *. Kết quả thực nghiệm của nghiên cứu cho thấy mẫu episode mở rộng chứa nhiều thông tin hỗ trợ mô hình dự đoán.*

** dữ liệu AIM-94 cung cấp bởi Michael Kahn, Washington University, St. Louis, MO.*

1 GIỚI THIỆU

Khai khoáng dữ liệu có phụ thuộc vào thời gian là một trong mười vấn đề lớn của ngành khai khoáng dữ liệu hiện nay [15]. Việc dự đoán các sự kiện chưa xảy ra từ dữ liệu thời gian có tiềm năng ứng dụng rất lớn nhưng chưa được nghiên cứu đầy đủ. Trong các giải thuật khai khoáng mẫu tuần tự [1, 2, 3, 4, 6] đã đề ra mô hình khai khoáng *chuỗi dữ liệu con* phổ biến từ cơ sở dữ liệu. Những chuỗi dữ liệu con có mang tính chất thứ tự nhưng khó truy xuất thông tin nội tại giữa các thành tố bên trong chuỗi. Việc đánh giá kết quả dựa trên độ đo ủng hộ hay tần suất xuất hiện của chuỗi. Các giải thuật sinh ra chuỗi dữ liệu con có độ phức tạp cao, duyệt qua cơ sở dữ liệu nhiều lần nên tốn nhiều thời gian xử lý, sử dụng bộ nhớ lớn. Một phương hướng tiếp cận khác theo mô hình khai khoáng mẫu Episode [5] trình bày bởi Mannila và các đồng sự với mục tiêu tìm kiếm các episode phổ biến. Các episode này là một tập các sự kiện xuất hiện phổ biến cùng nhau trong chuỗi sự kiện. Các mẫu episode tuần tự và mẫu episode song song được sinh ra qua quá trình quét cạn toàn bộ chuỗi tuần tự. Các mô hình của tác giả đề xuất tách biệt hoàn toàn 2 loại mẫu tuần tự và mẫu song song với nhau, nhưng trong dữ liệu thực tế cho thấy 2 loại mẫu này có quan hệ mật thiết với nhau (theo mối quan hệ nguyên nhân – kết quả). Thông tin của các mẫu episode bao gồm số lần xuất hiện của mẫu cùng với mối quan hệ giữa các sự kiện trong mẫu. Tác giả có định nghĩa luật tuần tự là mối quan hệ giữa mẫu episode và mẫu *episode con,* tiêu chí để đánh giá luật tuần tự là độ tin cậy của luật và được tính

dựa trên độ ủng hộ của mẫu episode với mẫu episode con. Trong bài viết của Diệp và các đồng sự [9], tác giả đề xuất giải thuật khai khoáng mẫu episode hiệu quả hơn với lượng mẫu phổ biến nhiều hơn, thời gian xử lý nhanh và các luật tuần tự thu được có độ tin cậy cao. Giải thuật tập trung khai khoáng các mẫu episode trên những sự kiện mà người sử dụng quan tâm và các sự kiện này thật sự xuất hiện trên chuỗi dữ liệu tuần tự. Tuy lượng thông tin thu được lớn nhưng do mặt hạn chế của mô hình mẫu episode nên kết quả thu được khó ứng dụng vào trong thực tế. Trong quá trình khai khoáng mẫu episode có những episode có những thứ tự đặc biệt được Katoh gọi theo đặc trưng riêng như sectorial episode, diamond episode, elliptic episode [10, 11, 14]. Trong bài viết này chúng tôi sẽ đề xuất kết hợp mẫu tuần tự và mẫu song song episode tạo ra extended episode, việc đánh giá mẫu extended episode dựa trên độ đo ủng hộ thuần và tần suất xuất hiện thuần của mẫu.

Phần tiếp theo của bài viết này được trình bày theo thứ tự như sau: phần 2 trình bày về mô hình extended episode. Phần 3 trình bày về giải thuật khai khoáng extended episode chúng tôi đề xuất. Kết quả thực nghiệm được trình bày trong phần 4. Phần 5 là kết luận và hướng phát triển.

2 MÔ HÌNH MẪU EXTENDED EPISODE

Giả sử mỗi sự kiện xảy ra ở một thời điểm t xác định là một số tự nhiên. Cho một tập E là tập hợp các sự kiện. Một sự kiện (e, t) là một sự kiện khi và chỉ khi $e \in \mathsf{E}$ và t là một số nguyên dương chỉ thời gian xảy ra sự kiện này. Sự kiện e có thể chứa một hoặc nhiều thuộc tính nhưng để đơn giản cho việc trình bày chúng ta giả định mỗi sự kiện chỉ chứa một thuộc tính. Theo đó ta có tập sự kiện trên chuỗi sự kiện :{ (e, t) | $e \in \mathsf{E}$ } với t > 0.

Định nghĩa 1. Một chuỗi tuần tự S trên tập E là một bộ 3(S, T_s, T_e) trong đó T_s và T_e là số nguyên dương chỉ thời gian trên S. T_s là thời gian bắt đầu, T_e là thời gian kết thúc, $T_e - T_s = l_s$ và $T_s \le t_i \le T_e$ với mọi i = 1,…,n và S={ (A_1,t_1) , (A_2,t_2) , (A_3,t_3) , …. , (A_n,t_n) }là một chuỗi tuần tự các sự kiện mà $A_i \in E$ $\forall$ i = 1,…,n và $t_i \le t_{i+1}$ $\forall$ i=1,….,n-1. Định nghĩa này có một chút khác biệt với định nghĩa chuỗi tuần tự S' trong [4, 5] với S'={ (A'_1,t'_1) , (A'_2,t'_2) , (A'_3,t'_3) , …. , (A'_n,t'_n) } $A'_i \in E$ $\forall$ i = 1,…,n và $t'_i < t'_{i+1}$ $\forall$ i=1,….,n-1. Như vậy, S có thể chứa nhiều sự kiện xảy ra tại một thời điểm t so với S' chỉ có thể chứa tối đa 1 sự kiện tại một thời điểm t'.

Định nghĩa 2. Một mẫu (Episode) được xét là một bộ 3 (V, ≤ ,g) trong đó V là tập hợp của các nút, ≤ là một quan hệ thứ tự bán phần trên V và ánh xạ g: V $\rightarrow$ E là một ánh xạ kết hợp giữa các nút với các loại sự kiện. Kích thước của α kí hiệu |α| và kích thước V kí hiệu |V|. Episode α là tuần tự khi mối quan hệ ≤ là có thứ tự, Episode α là song song khi mối quan hệ ≤ là tầm thường.

Định nghĩa 3. Tần suất của mẫu X *freqX* trên chuỗi S với thời gian t_bound có công thức như sau:

$$freq(X) = \frac{\sup port(X)_{t_bound}}{\frac{Te - Ts}{t_bound}}$$

Trong đó freq(X) là tần suất xuất hiện của mẫu X trên chuỗi S với khoảng thời gian t_bound, support(X) $_{t_bound}$ là độ đo ủng hộ của mẫu X trên S với khoảng thời gian t_bound.

Định nghĩa 4. Gọi σ là ngưỡng độ ủng hộ nhỏ nhất với 0 < σ < 1. Khi đó một mẫu X là mẫu phổ biến khi và chỉ khi support (X) ≥ σ.

Định nghĩa 5. Luật episode là biểu thức $\beta \Rightarrow \gamma$, với β và γ là các episodes sao cho β là episode con của γ. Episode β là episode con của γ ($\beta \prec \gamma$), nếu đồ thị biểu diễn β là đồ thị con của đồ thị biểu diễn γ. Độ tin cậy của episode γ được tính: confγ = fr(γ, S, t_bound)/ fr(β, S, t_bound)

Định nghĩa 6. Gọi λ là ngưỡng độ tin cậy nhỏ nhất với 0 < λ < 1. Khi đó luật tuần tự X là luật tin cậy khi và chỉ khi conf (X) ≥ λ.

Định nghĩa 7. Độ ủng hộ thuần của 1 mẫu episode [B|A] là độ ủng hộ episode [B|A] và không tồn tại bất kỳ episode [BX|A] ($\forall$ X ≠ B, X và B cùng xảy ra trước A) rsupport [B|A] = support [B|A] – support [B$\cap$X|A].

Định nghĩa 8. Mẫu mở rộng episode có dạng E→B, với E là mẫu episode song song, B là mẫu episode tuần tự.

3 GIẢI THUẬT KHAI KHOÁNG EXTENDED EPISODE

Việc khai khoáng các mẫu extended episode được thực hiện qua 3 giai đoạn

3.1 Giai đoạn 1 - Tạo ra các mẫu episode kèm theo các thông tin

Giải thuật IniE được sử dụng để trích xuất các mẫu episode từ chuỗi dữ liệu tuần tự đưa vào. Giải thuật được mô tả qua ngôn ngữ giả bên dưới.

Thủ tục IniE (E; t_bound; S)

Đầu vào: chuỗi tuần tự S = ⟨S1; : : : Sn⟩ với thời gian xảy ra tương ứng của mỗi sự kiện; thời gian giới hạn t_bound, Vector danh sách sự kiện E =∅, mỗi phần tử trong E chứa 1 vector lưu giá trị thời gian xuất hiện các sự kiện $E_i = \emptyset$.

Đầu ra: Vector C chứa các mẫu tìm được với thông tin thời điểm xuất hiện.

Nội dung giải thuật:

```
1  C := ∅;
2  E := ∅;
3  foreach(s ∈ S)
4     if (s ∉ E) then
5        E := E ∪ {s}
6     Es := Es ∪ {t(s)}
7  end(foreach)
8  foreach(e ∈ E)
9     foreach(α # e and α ∈ E )
10       if ([e|α] ∉ C) then
11          C:= C ∪ {[e|α]}
12       foreach(i∈Ee, j∈Eα and t(Eα[j])-t(Ee[i])
   < t_bound)
13       C[e|α] := C[e|α] ∪ [t(Ee[i]), t(Eα[j])]
14       end (foreach)
15    end (foreach)
16 end (foreach)
17 return C;
```

Giải thuật IniE duyệt qua chuỗi dữ liệu S duy nhất 1 lần, mỗi khi sự kiện xuất hiện sẽ được lưu vào vector E cùng với thời điểm xuất hiện của sự kiện. Sau khi duyệt qua hết chuỗi dữ liệu S, vector E chứa toàn bộ các sự kiện kèm thời điểm xuất hiện của các sự kiện. Lần lượt xét các sự kiện trong E, với mỗi cặp sự kiện khác nhau có thời điểm xuất hiện thỏa mãn tham số t_bound sẽ tạo mẫu tuần tự episode tương ứng với hai sự kiện khác nhau đó, thông tin về thời điểm xuất hiện cũng sẽ được lưu vào vector C. Vòng lặp kết thúc khi không tìm được thêm các cặp sự kiện thỏa mãn yêu cầu. Kết thúc giai đoạn 1 ta thu được vector C chứa danh sách các mẫu episode và thời điểm xuất hiện của từng mẫu episode, vector E chứa các sự kiện và thời điểm xuất hiện của các sự kiện.

Định lý 1: Độ phức tạp của IniE là $O(n+ l^2m)$.

Chứng minh: Với vòng lặp đầu tiên duyệt qua chuỗi tuần tự S để lưu lại thông tin và thời điểm xảy ra sự kiện có độ phức tạp tuyến tính phụ thuộc vào độ lớn n của chuỗi sự kiện S. Trong vòng lặp, mỗi thời điểm sự kiện s xuất hiện trong S sẽ được lưu thời gian lại theo loại trong vector sự kiện (E) nên thời gian tính toán sẽ tuyến tính theo **n**. Vậy độ phức tạp từ dòng 1-7 là O (n). Vòng lặp thứ 2 từ dòng 8 đến dòng 16 lặp qua từng loại sự kiện trong vector sự kiện E. Vòng lặp này chứa vòng lặp con lặp lại qua từng sự kiện trong E và không xét 2 sự kiện trùng nhau nên độ phức tạp lúc này là l^2 (với l là số sự kiện trong E). Với mỗi mẫu tuần tự episode được tạo ra, lần lượt xét thời gian từng cặp sự kiện tạo nên mẫu với vòng lặp từ dòng 11 đến dòng 14. Vòng lặp này sẽ lặp qua **m** lần thời gian của sự kiện xuất hiện nhiều (α hoặc e). Mỗi cặp thời gian thỏa ràng buộc tổng thời gian nhỏ hơn độ lớn t_bound sẽ được thêm vào vector C tương ứng với mẫu được tạo ra bởi 2 sự kiện đó. Như vậy độ phức tạp thủ tục IniE là $O (n + l^2m)$. Trong thực tế, số lần xuất hiện **m** của các sự kiện và số sự kiện l trong E nhỏ hơn độ lớn **n** của chuỗi dữ liệu S nhiều lần nên giải thuật IniE là chấp nhận được.

3.2 Giai đoạn 2 – Tạo các mẫu Episode dạng mở rộng và hiệu chỉnh thông tin trên các mẫu

Giải thuật ExtEE đọc thông tin các mẫu episode trong vector C và thông tin các sự kiện trong vector E tạo ra ở giai đoạn 1. Đầu tiên khởi tạo vector M để chứa các mẫu mở rộng và thông tin độ đo ủng hộ thuần. Lần lượt xét các cặp mẫu trong vector C, với mỗi cặp mẫu có cùng sự kiện cuối lần lượt xét thời điểm xuất hiện của cặp mẫu. Nếu mẫu mở rộng chưa có trong vector M thì thêm vào vector M. Mỗi cặp thời điểm thỏa giá trị t_bound sẽ được lưu lại trong vector M theo tương ứng với từng mẫu trong M và cập nhật độ đo ủng hộ của các mẫu trong C. Các cặp thời điểm thỏa giá trị t_bound sẽ có 3 trường hợp xảy ra: 2 mẫu cùng xuất hiện đúng vị trí trong thời gian t_bound, mẫu thứ nhất có thời điểm bắt đầu và thời điểm kết thúc nhỏ hơn thời điểm mở đầu và kết thúc của mẫu thứ hai (hoặc ngược lại), mẫu thứ nhất được chứa trong mẫu thứ hai (hoặc ngược lại). Kết thúc giải thuật, vector M sẽ chứa các mẫu mở rộng với độ đo ủng hộ thuần. Giải thuật được mô tả chi tiết bằng ngôn ngữ giả như sau:

Thủ tục ExtEE (E; C; t_bound)
Đầu vào: Vector C chứa các mẫu Episode tuần tự, Vector E chứa danh sách sự kiện thu được ở giai đoạn 1, thời gian giới hạn t_bound,
Đầu ra: Vector M chứa các mẫu mở rộng với độ ủng hộ thuần của từng mẫu;
Giải thuật:

```
1  M := ∅;
2  foreach(A[α|x] ∈ C, B[β|x] ∈ C and α ≠ β )
3     M:= M ∪ {[αβ|x]}
4     foreach([i1,i2]∈A[α|x], [j1,j2]∈Eβ)
5        if ([i1,i2]⊂ [j1,j2]) then
6           M[αβ|x] := M[αβ|x] ∪ [j1,j2]
7           C[α|x]:= C[α|x] – [i1,i2]
8           C[β|x]:= C[β|x] – [j1,j2]
9        if([i1,i2]⊃ [j1,j2]) then
10          M[αβ|x] := M[αβ|x] ∪ [i1,i2]
11          C[α|x]:= C[α|x] – [i1,i2]
12          C[β|x]:= C[β|x] – [j1,j2]
13    end(foreach)
14    if(M[αβ|x] = ∅)
15       M:= M - {[αβ|x]}
16 end(foreach)
17 return M;
```

Định lý: Độ phức tạp của ExtEE là O (n.m)

Chứng minh: ExtEE thực hiện chủ yếu với vòng lặp từ dòng 2 đến dòng 16. Vòng lặp này sẽ lặp qua **n** cặp mẫu tuần tự episode có chứa cùng 1 loại sự kiện X ở thứ tự cuối cùng của mẫu trong vector chứa mẫu C (vector mẫu tạo được ở giai đoạn 1). Vòng lặp này chứa một vòng lặp con từ dòng 5 đến dòng 13, vòng lặp bên trong này sẽ lặp qua **m** các cặp thời gian của mẫu episode A và B. Khi các cặp thời gian này thỏa mẫu A và B cùng xuất hiện trong khung thời gian và khung thời gian này nhỏ hơn ngưỡng t_bound thì sẽ tạo mẫu mở rộng có dạng [(AB) →X]. Nếu M không chứa mẫu dạng này thì sẽ thêm vào M. Nếu M đã chứa mẫu thì thêm khoảng thời gian lớn nhất chứa cả 2 mẫu tuần tự. Độ phức tạp của 2 vòng lặp này là O (n.m).

3.3 Giai đoạn 3 – Kết hợp thông tin để dự đoán

Sử dụng vector M chứa các mẫu mở rộng cùng với thông tin độ đo ủng hộ thuần thu được từ giai đoạn 2 để sinh ra các luật tuần tự thuần. Việc dự đoán chủ yếu dựa trên luật tuần tự thuần với công thức như sau:

$$p(Y) = \sum [X \mid Y]_{rconf} \ (\forall\, T_S\text{-}[X|Y]_{ts} \leq t_bound)$$

Trong đó T_S là thời điểm kết thúc của chuỗi dữ liệu sự kiện, ts là thời điểm xuất hiện cuối cùng của mẫu có dạng [X|Y] với Y là một sự kiện bất kỳ (Y∈E), rconf là độ tin cậy thuần của mẫu [X|Y].

4 KẾT QUẢ THỰC NGHIỆM

Áp dụng các giải thuật khai khoáng mẫu extended episode vào cơ sở dữ liệu tiểu đường AIM-94 cung cấp bởi Michael Kahn, Washington University, St. Louis, MO. Dữ liệu bao gồm 70 chuỗi dữ liệu tuần tự (tương ứng với 70 bệnh nhân). Mỗi chuỗi trung bình có 500 mốc thời gian với 700 sự kiện xảy ra. Dữ liệu có 20 loại sự kiện bao gồm 12 loại là dữ liệu liên tục và 8 loại là dữ liệu rời rạc. Trong 12 loại dữ liệu liên tục có 3 loại là các liều insulin và 9 loại còn lại là chỉ số đường trong máu. Đơn vị thời gian được tính là giờ, khoảng thời gian đo thông thường là các giờ: ăn sáng (08:00), ăn trưa (12:00), ăn chiều (18:00) và ăn khuya (22:00). Sau khi chuẩn hóa dữ liệu chúng tôi thu được các loại sự kiện sau:

Có 9 loại sự kiện về lượng đường trong máu được minh họa chi tiết trong Bảng 1:

Bảng 1: Bảng rời rạc dữ liệu đường trong máu

Chỉ số đường trong máu	Sự kiện
Tăng trên 25%	Tăng rất mạnh
Tăng 17% → 25%	Tăng mạnh
Tăng 12% → 17%	Tăng vừa
Tăng 5% → 12 %	Tăng nhẹ
Dao động 5% → 5%	Không đổi
Giảm 5% → 12%	Giảm nhẹ
Giảm 12 → 17%	Giảm vừa
Giảm 17 → 25%	Giảm mạnh
Giảm hơn 25%	Giảm rất mạnh

Có 9 loại sự kiện về liều lượng insulin được minh họa trên các Bảng 2, 3, 4.

Bảng 2: Bảng rời rạc dữ liệu Regular insulin

Lượng insulin tác dụng nhanh (Regular insulin dose)	Sự kiện
Dưới 6 đơn vị	Liều R1
Từ 6 đến 12 đơn vị	Liều R2
Trên 12 đơn vị	Liều R3

Bảng 3: Bảng rời rạc dữ liệu NHP insulin

Lượng insulin tác dụng trung gian (NHP insulin dose)	Sự kiện
Dưới 12 đơn vị	Liều NHP1
Từ 12 đến 20 đơn vị	Liều NHP2
Trên 20 đơn vị	Liều NHP3

Bảng 4: Bảng rời rạc dữ liệu UltraLente insulin

Lượng insulin tác dụng chậm (UltraLente insulin dose)	Sự kiện
Dưới 7 đơn vị	Liều U1
Từ 7 đến 13 đơn vị	Liều U2
Trên 13 đơn vị	Liều U3

Ngoài ra theo quan điểm y học, chúng tôi chọn từ cơ sở dữ liệu 13 sự kiện sau: trước khi ăn sáng, sau khi ăn sáng, trước khi ăn trưa, sau khi ăn trưa, trước khi ăn vặt, có triệu chứng hạ đường trong máu, bữa ăn kiêng thông thường, bữa ăn kiêng nhiều hơn thông thường, bữa ăn kiêng ít hơn thông thường, hoạt động thể dục thông thường, hoạt động thể dục nhiều hơn thông thường, hoạt động thể dục ít hơn thông thường, sự kiện đặc biệt (tiệc, lễ hội...). Tổng hợp lại có 31 loại sự kiện.

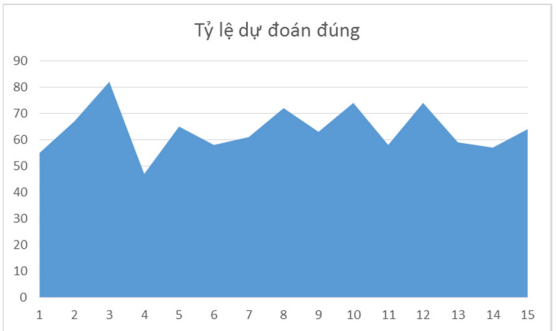
Các thông số cho chương trình được xác định: ràng buộc thời gian t_bound = 24 (giờ), ngưỡng ủng hộ nhỏ nhất $support_{min}$ = 0,2 %, ngưỡng ủng hộ thuần $rsupport_{min}$ = 0,1%. Giải thuật được cài đặt trên hệ điều hành Window, viết bằng ngôn ngữ lập trình JAVA với công cụ phát triển NETBEAN 7.0.

Bảng 5: Kết quả áp dụng các giải thuật và độ đo khác nhau trên tập dữ liệu AIM-94

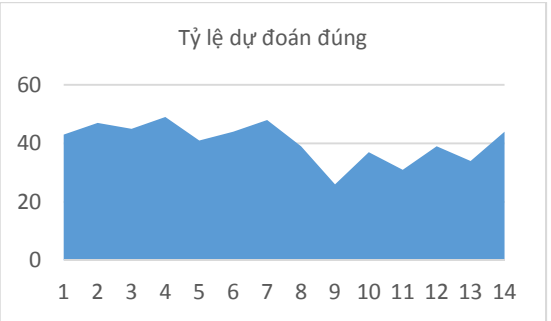
Độ đo	Episode	Frequent Episode	Extended Episode	Frequent Extended Episode
support	823	672	1356	1219
rsupport	821	631	1352	936

Bảng 6: Thời gian thực thi của các giải thuật trên các chuỗi dữ liệu khác nhau

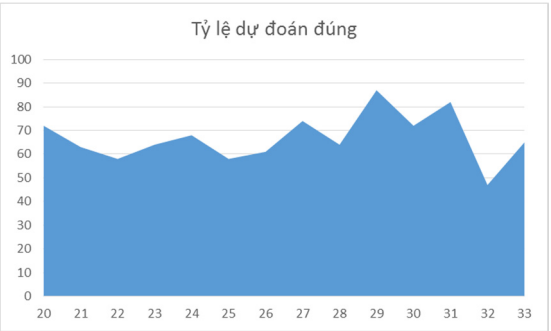
Thời gian thực hiện (ms)	Data-01	Data-12	Data-39	Thời gian Trung bình
WINEPI	823	672	645	713
MINEPI	730	631	589	650
DYNEPI	525	487	471	494
IniE + ExtEE	780	656	612	682



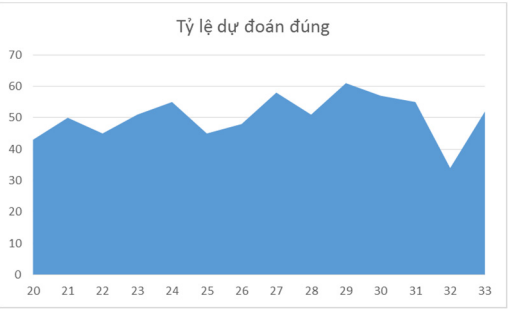
Hình 1: Trung bình kết quả dự đoán trên từng chuỗi dữ liệu sử dụng độ đo ủng hộ thuần, mẫu extended episode và luật tuần tự thuần (đơn vị:%)



Hình 2: Trung bình kết quả dự đoán trên từng chuỗi dữ liệu sử dụng độ đo ủng hộ, mẫu episode và luật tuần tự (đơn vị:%)



Hình 3: Trung bình kết quả dự đoán trên từng chuỗi dữ liệu sử dụng độ đo ủng hộ thuần, mẫu extended episode và luật tuần tự thuần (đơn vị:%)



Hình 4: Trung bình kết quả dự đoán trên từng chuỗi dữ liệu sử dụng độ đo ủng hộ, mẫu extended episode và luật tuần tự (đơn vị:%)

Bảng so sánh ở Bảng 5 cho thấy, khi sử dụng độ đo ủng hộ thuần (rsupport), tổng số mẫu episode và extended episode luôn thấp hơn khi sử dụng độ đo ủng hộ (support). Vì độ đo ủng hộ thuần xác định chính xác số lần mẫu xuất hiện trong khoảng thời gian t_bound giúp tránh trường hợp 1 mẫu được xét ủng hộ 2 lần. Thông tin dự đoán chủ yếu dựa trên luật tuần tự thuần, mà luật tuần tự thuần phụ thuộc chủ yếu vào độ ủng hộ thuần của mẫu. Do đó kết quả dự đoán sử dụng độ đo ủng hộ thuần

và luật tuần tự thuần cho kết quả chính xác hơn (Hình 3 và Hình 4).

Hình 2 cho thấy việc sử dụng mẫu episode để dự đoán cho kết quả có độ chính xác thấp nhất trong cả 4 trường hợp.

Bảng 6 cho thấy tổng thời gian thực hiện của giải thuật cả 2 giải thuật IniE và ExtEE tốt hơn giải thuật WINEPI

5 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi vừa trình bày mô hình mẫu episode mở rộng (extended episode) thay thế cho mẫu episode trong các ứng dụng dự đoán dựa trên dữ liệu thời gian. Trong thực tế với trường hợp dữ liệu đa sự kiện (các sự kiện xảy ra cùng một thời điểm) thì không thể sử dụng mẫu episode được. Ngoài ra chúng tôi còn đề xuất độ đo ủng hộ thuần, độ tin cậy thuần thay thế cho độ đo ủng hộ và độ tin cậy. Kết quả thực nghiệm cho thấy các cải tiến trên hỗ trợ tốt hơn cho việc dự đoán sự kiện tương lai. Trong thời gian tới, nhóm chúng tôi sẽ cải tiến mô hình hoạt động linh hoạt hơn có thể tận dụng được thông tin dữ liệu mẫu episode có sẵn và có thể phát hiện các sự kiện *ẩn* trong chuỗi dữ liệu.

TÀI LIỆU THAM KHẢO

1. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo: Fast discovery of association rules, in U. M. Fayyed, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (eds.): Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 307–328, 1996.
2. R. Agrawal, R. Srikant: Fast algorithms for mining association rules in large databases, Proc. 20th VLDB, 487–499, 1994.
3. R. Agrawal, R. Srikant: Mining sequential patterns, Proc. 11th ICDE, 3–14, 1995.
4. C. Bettini, S. Wang, S. Jajodia, J.-L. Lin: Discovering frequent event patterns with multiple granularities in time sequences, IEEE Trans. Knowledge and Data Engineering10, 222–237, 1998.
5. H. Mannila, H. Toivonen, A. I. Verkamo: Discovery of frequent episodes in event sequences, Data Mining and Knowledge Discovery 1, 259–289, 1997.
6. J. Pei, J. Han, B. Mortazavi-Asi, J. Wang, H. Pinto, Q. Chen, U. Dayal, M.-C. Hsu: Mining sequential patterns by pattern-growth: The PrefixSpan approach, IEEE Trans. Knowledge and Data Engineering16, 1–17, 2004.
7. R. Srikant, R. Agrawal: Mining sequential patterns: Generalizations and perfor-mance improvements, Proc. 5th EDBT, 3–17, 1996..
8. S. Tsumoto: Guide to the bacteriological examination data set, in E. Suzuki (ed.): Proc. International Workshop of KDD Challenge on Real-World Data (KDD Challenge 2000), 8–12, 2000.
9. Ba-Diep Nguyen, Xuan-Hiep Huynh, Julien Blanchard : Phát hiện mẫu tuần tự với kích thước thay đổi bằng giải thuật DYNEPI, Kỷ yếu Hội nghị khoa học 5 năm nghiên cứu khoa học Khoa CNTT Trường Đại học Cần Thơ, 2011.
10. Katoh, T., Hirata, K., Harao, M.: Mining frequent diamond episodes from event sequences. In: Torra, V., Narukawa, Y., Yoshida, Y. (eds.) MDAI 2007. LNCS (LNAI), vol. 4617, pp. 477–488. Springer, Heidelberg (2007).
11. Katoh, T., Hirata, K., Harao, M., Yokoyama, S., Matsuoka, K.: Extraction of sectorial episodes representing changes for drug resistant and replacements of bacteria. In: Proc. CME 2007, pp. 304–309 (2007).
12. Katoh, T., Arimura, H., Hirata, K.: A polynomial-delay polynomial-space algorithm for extracting frequent diamond episodes from event sequences. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS (LNAI), vol. 5476, pp. 172–183. Springer, Heidelberg (2009).
13. Katoh, T., Hirata, K.: A simple characterization on serially constructible episodes. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 600–607. Springer, Heidelberg (2008).
14. Katoh, T., Hirata, K.: Mining frequent elliptic episodes from event sequences. In: Proc. 5th LLLL, pp. 46–52 (2007).
15. 1Q. Yang and X. Wu. 10 Challenging Problems in Data Mining Research. Journal of Information Technology & Decision Making 5(4):597-604, 2006.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

XÂY DỰNG MẠNG CẢM BIẾN KHÔNG DÂY TRONG NÔNG NGHIỆP CHÍNH XÁC

Lê Đình Tuấn[1] và Thái Doãn Ngọc[1]

[1] *Trường Đại học Kinh tế Công nghiệp Long An*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Wireless sensor networks in agriculture

Từ khóa:
Mạng cảm biến không dây, nông nghiệp chính xác

Keywords:
Wireless sensor networks, precision agriculture

ABSTRACT

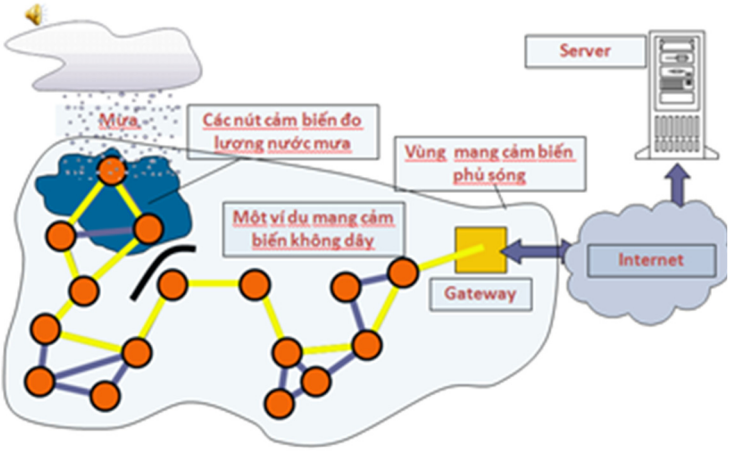
This paper describes the design, implementation, and deployment of wireless sensor network for precision agriculture at Long An University of Economics and Industry. Precision agriculture is a farming management concept which applies the right amount of input (water, fertilizer, pesticide, etc.) at the right location and at the right time to enhance production and improve quality, while protecting the environment. Wireless sensor network is built for monitoring and controlling environmental parameters, such as the environmental parameters (air temperature, air humidity, light, insects, weeds, etc.) and other parameters related to soil conditions (soil moisture, pH, etc.). The data are collected, stored and transmitted wirelessly to the farmers to handle, through which they can control and take appropriate actions for their farm to increase production and quality. From the hardware side, the system consists of three components: wireless sensor nodes DHLA-WSN, wireless sensor node manager DHLA-WMN, and a server controller. Software was built at each node to carry out their task. The deployed system is testing on the field, working properly and promising which brings significantly benefits to the agriculture field.

TÓM TẮT

Bài báo này mô tả việc thiết kế, xây dựng và triển khai mạng cảm biến không dây phục vụ nông nghiệp chính xác tại Trường Đại học Kinh tế Công nghiệp Long An. Nông nghiệp chính xác hoặc canh tác chính xác là kỹ thuật áp dụng đúng số lượng đầu vào (nước, phân bón, thuốc trừ sâu,...) vào đúng vị trí và vào đúng thời điểm để tăng cường sản xuất và nâng cao chất lượng [1]. Mạng cảm biến không dây được xây dựng nhằm phục vụ yêu cầu của nông nghiệp chính xác. Thông qua mạng cảm biến không dây, người nông dân có thể thu thập được các thông số đến quy trình sản xuất nông nghiệp như nhiệt độ không khí, độ ẩm không khí, ánh sáng, độ ẩm đất, độ pH,... Các thông tin này được thu thập, lưu trữ và truyền tải không dây đến trạm quản lý để phân tích và xử lý, qua đó người sử dụng có thể điều khiển và đưa ra các quyết định phù hợp nhằm tăng cường sản xuất và nâng cao chất lượng cây trồng. Hệ thống được thiết kế và xây dựng cả phần cứng và phần mềm. Về phần cứng, hệ thống gồm 3 phần chính: các nút cảm biến DHLA-WSN tại khu vực giám sát, các nút quản lý vùng DHLA-WMN, và trung tâm điều hành (server). Tại mỗi nút các phần mềm được xây dựng để lập trình hệ thống hoạt động theo yêu cầu. Hệ thống đã chạy thử nghiệm tại trường bước đầu hoạt động tốt và hứa hẹn đem lại nhiều lợi ích đáng kể trong hệ thống sản xuất nông nghiệp.

1 GIỚI THIỆU MẠNG CẢM BIẾN KHÔNG DÂY

Mạng cảm biến không dây (Wireless Sensor Networks, gọi tắt là WSNs) là một mạng bao gồm nhiều nút cảm biến (sensor nodes), được trang bị các nút cảm biến có khả năng cảm biến môi trường như cảm biến nhiệt độ, cảm biến độ ẩm, cảm biến cường độ ánh sáng,... và có khả năng giao tiếp không dây (wirelessly) với các nút còn lại tạo thành một mạng cảm biến không dây phủ sóng một vùng vật lý nào đó nhằm giám sát, theo dõi và quản lý vùng đó [2]. Thông tin thu được từ một nút cảm biến sẽ được truyền về một trạm gốc (base station hay gateway) thông qua các nút cảm biến khác, và cuối cùng thông qua Internet truyền về trung tâm dữ liệu để lưu trữ, phân tích và xử lý.



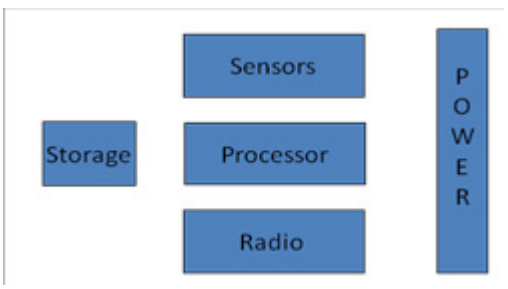
Hình 1: Một ví dụ về mạng cảm biến không dây theo dõi lượng nước mưa

Một nút cảm biến thông thường bao gồm các thành phần sau: Bộ vi xử lý nhỏ và sử dụng năng lượng ít (low power processor); Bộ nhớ (Memory); Radio để truyền dữ liệu không dây; nguồn điện/pin (power source); và các bộ cảm biến (sensors).

Mạng cảm biến không dây được ứng dụng rộng rãi trong nhiều lĩnh vực như: giám sát kết cấu của công trình (cầu, tòa nhà...); theo dõi cuộc sống của các loài thú hoang dã; cảnh báo cháy rừng; phát hiện rò rỉ hóa chất trong các nhà máy; giám sát các tòa nhà thông minh…

2 ỨNG DỤNG MẠNG CẢM BIẾN KHÔNG DÂY TRONG NÔNG NGHIỆP

Theo truyền thống nông nghiệp là việc thực hiện một nhiệm vụ cụ thể, chẳng hạn như trồng hoặc thu hoạch, với một lịch trình định trước. Nhưng bằng cách thu thập dữ liệu thời gian thực về thời tiết, đất và chất lượng không khí, theo dõi sự trưởng thành của cây trồng và thậm chí cả trang thiết bị và chi phí lao động, các phân tích có thể được sử dụng để đưa ra quyết định thông minh hơn. Đây được gọi là nông nghiệp chính xác (hoặc canh tác chính xác). Một định nghĩa của nông nghiệp chính xác có thể là như sau: kỹ thuật áp dụng đúng số lượng đầu vào (nước, phân bón, thuốc trừ sâu,…) vào đúng vị trí và vào đúng thời điểm để tăng cường sản xuất và nâng cao chất lượng [1].



Hình 2: Cấu trúc một nút cảm biến

Với nông nghiệp chính xác, trung tâm kiểm soát thu thập và xử lý dữ liệu trong thời gian thực để giúp nông dân đưa ra quyết định tốt nhất liên quan đến trồng, bón phân và thu hoạch cây trồng có. Các nút cảm biến được đặt tại nơi trồng để đo nhiệt độ và độ ẩm của đất và không khí xung quanh. Ứng dụng cảm biến không dây trong nông nghiệp chính xác nâng cao hiệu quả, năng suất và lợi nhuận trong nhiều hệ thống sản xuất nông nghiệp, trong khi giảm thiểu tác động không mong muốn đến địa điểm nơi trồng. Các thông tin thời gian thực thu được từ các lĩnh vực có thể cung cấp một cơ sở vững chắc cho nông dân để điều chỉnh chiến lược bất cứ lúc nào. Thay vì đưa ra quyết định dựa vào một số điều kiện trung bình giả thuyết hay kinh nghiệm chủ quan của cá nhân, có thể không tồn tại bất cứ nơi nào trong thực tế, một cách tiếp cận nông nghiệp chính xác nhận ra sự khác biệt và điều chỉnh hoạt động quản lý tối ưu.

Việc ứng dụng mạng cảm biến không dây để quản lý hoạt động nông nghiệp chính xác làm giảm đáng kể số lượng đầu vào như phân bón, nước, thuốc trừ sâu... được sử dụng trong khi tăng sản lượng. Nông dân do đó thu được lợi nhuận trên đầu tư của mình bằng cách tiết kiệm chi phí kiểm dịch thực vật và phân bón. Áp dụng đúng số lượng đầu vào ở đúng nơi và đúng thời điểm đem lại lợi ích cho cây trồng, trong khi tiết kiệm nguồn tài nguyên như đất và nước ngầm, và do đó tối ưu hóa toàn bộ chu kỳ trồng trọt. Nông nghiệp bền vững tìm cách để đảm bảo một nguồn cung cấp liên tục của thực phẩm trong giới hạn sinh thái, kinh tế và xã hội cần thiết để duy trì sản xuất trong dài hạn. Do đó độ chính xác nông nghiệp bằng cách sử dụng mạng cảm biến không dây sẽ cho phép theo đuổi mục tiêu này.

Ngoài ra, mạng cảm biến không dây còn được sử dụng để điều khiển trong nhà kính [3][4][5]. Một nhà kính là một cấu trúc bao phủ mặt đất thường được sử dụng cho sự tăng trưởng và phát triển của cây. Cấu trúc này được gắn với mục đích bảo vệ cây trồng và cho phép một môi trường tốt hơn để phát triển. Sự bảo vệ này đủ để đảm bảo đem lại chất lượng cao trong sản xuất cây trồng. Chức năng chính của một nhà kính là để cung cấp một môi trường thuận lợi hơn so với bên ngoài. Các yếu tố chính liên quan đến hệ thống điều khiển nhà kính: nhiệt độ, độ ẩm, khí CO_2, nồng độ, bức xạ, nước và chất dinh dưỡng có thể được điều khiển bằng mạng cảm biến không dây. Hệ thống tưới tiêu thông minh sử dụng mạng cảm biến không dây cũng là một trong các nghiên cứu được quan tâm nhiều vì lợi thế trong việc tiết kiệm nhân lực và tiết kiệm nước [6][7][8]. Cây trồng cần ánh sáng mặt trời, các chất dinh dưỡng và nước để phát triển. Tất cả các nhà sản xuất nông nghiệp có một yêu cầu lượng nước tối thiểu hàng năm để tồn tại và yêu cầu lượng nước tối ưu hàng năm cho sản xuất tối đa. Do đó hệ thống tưới tiêu thông minh là cần thiết để nâng cao năng suất cây trồng.

3 XÂY DỰNG MẠNG CẢM BIẾN KHÔNG DÂY PHỤC VỤ NÔNG NGHIỆP TẠI TRƯỜNG ĐẠI HỌC KINH TẾ CÔNG NGHIỆP LONG AN

Phần này mô tả mô hình mạng cảm biến không dây phục vụ nông nghiệp chính xác và việc triển khai thử nghiệm tại Tường Đại học Kinh tế Công nghiệp Long An.

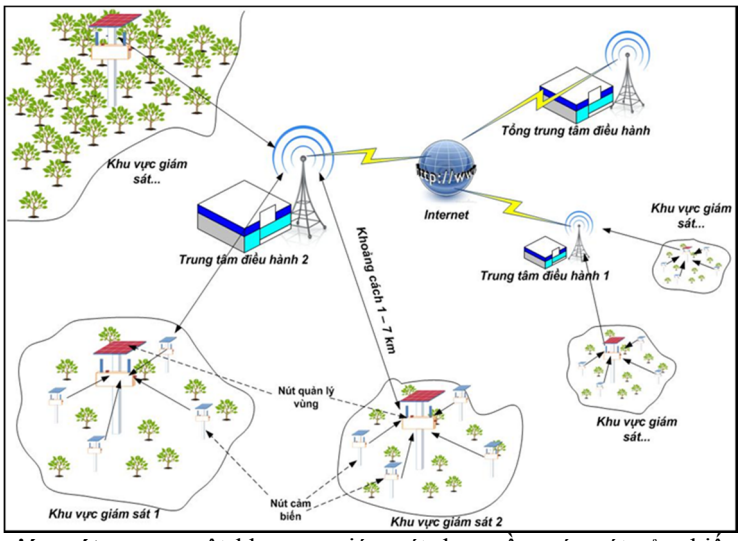
Hình 3: Một nút cảm biến được triển khai tại Trường Đại học KTCN Long An

3.1 Mô hình tổng thể của hệ thống

Hình 4 mô tả mô hình tổng thể của hệ thống mạng cảm biến không dây phục vụ nông nghiệp chính xác. Tại mỗi khu vực canh tác các nút cảm biến được triển khai nhằm theo dõi các thông số kỹ thuật liên quan đến chất lượng cây trồng như nhiệt độ không khí, độ ẩm không khí, ánh sáng,... và các thông số liên quan đến điều kiện đất trồng độ ẩm đất, độ pH,... Tại mỗi khu vực, các dữ liệu này được thu thập, lưu trữ và truyền tải không dây về một nút quản lý vùng, và sau đó dữ liệu được tiếp tục được truyền tải đến trung tâm điều hành và cuối cùng về trung tâm tổng điều hành thông qua mạng Internet. Dựa trên các thông số thu thập được, người dùng có thể xử lý và điều khiển từ các trung tâm điều hành hoặc trung tâm tổng điều hành để đưa ra các biện pháp thích hợp trong sản xuất nông nghiệp nhằm tăng cường sản xuất và nâng cao chất lượng. Ví dụ, nếu xác định độ ẩm đất tại một khu

vực thấp hơn yêu cầu, người dùng có thể ra lệnh từ xa kích hoạt máy bơm nước tại khu vực để cải thiện độ ẩm.

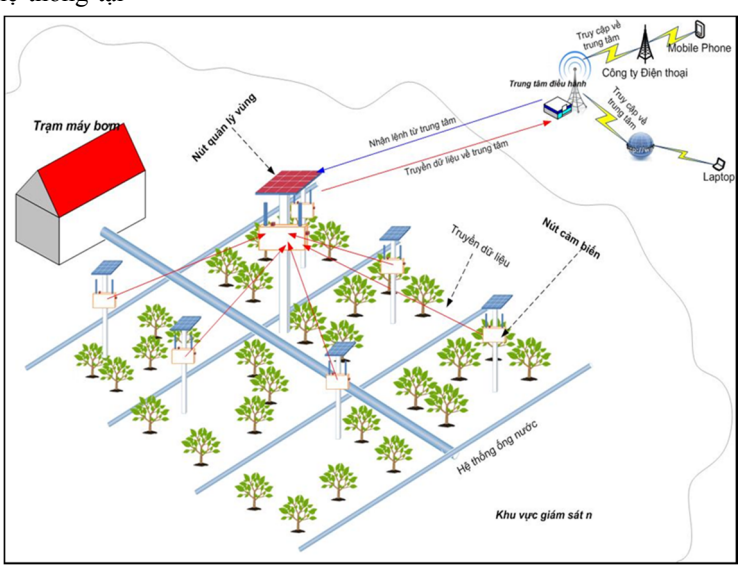
Do mô hình hệ thống có khả năng mở rộng nên hệ thống này có thể triển khai trên quy mô nhỏ (như một cánh đồng canh tác của một cá thể) đến quy mô lớn (như khu vực diện rộng ở cấp huyện, cấp thị xã, thành phố thậm chí cấp quốc gia).



Hình 4: Mô hình tổng thể của hệ thống

3.2 Mô hình chi tiết một khu vực giám sát

Hình 5 mô tả chi tiết cấu trúc của hệ thống tại một khu vực giám sát, bao gồm các nút cảm biến và một nút quản lý vùng.



Hình 5: Mô hình chi tiết một khu vực giám sát

a. Nút cảm biến (Sensor nodes)

Nút cảm biến (Sensor nodes) là một thiết bị được trang bị khả năng cảm biến môi trường như độ ẩm không khí, nhiệt độ, độ pH, lượng mưa, tốc độ và hướng gió, cường độ ánh sáng, hình ảnh... (tùy vào cảm biến được trang bị). Ngoài ra, nút cảm biến còn có khả năng giao tiếp với các nút cảm biến khác trong cùng 1 vùng phủ sóng và truyền tải không dây các dữ liệu thu thập được về trạm trung gian nội bộ.

b. Nút quản lý vùng (Field Managemnent Nodes)

Nút quản lý vùng là nút có nhiệm vụ quản lý các nút cảm biến trong vùng phủ sóng và đóng vai trò là nút trung gian giữa các nút cảm biến và trung tâm điều hành, nhận dữ liệu từ các nút cảm biến trong khu vực và truyền dữ liệu này về trung tâm điều hành gần nhất. Ngoài ra, nút quản lý vùng còn có nhiệm vụ nhận và truyền lệnh từ trung tâm điều

hành đến các nút cảm biến để thực thi lệnh theo yêu cầu của trung tâm điều hành.

3.3 Phần cứng

Phần này mô tả phần cứng của hệ thống được xây dựng tại khu vực giám sát và trung tâm điều hành.

3.3.1 Khu vực giám sát

a. Nút cảm biến (DHLA-WSN)

Hình 6 mô tả cấu trúc của một nút cảm biến trong hệ thống. Một nút cảm biến gồm 5 thành phần sau đây:

– Khối cung cấp năng lượng: khối này sẽ cấp nguồn điện 12V cho thiết bị hoạt động. Điện có thể lấy từ năng lượng mặt trời, pin dự trữ …

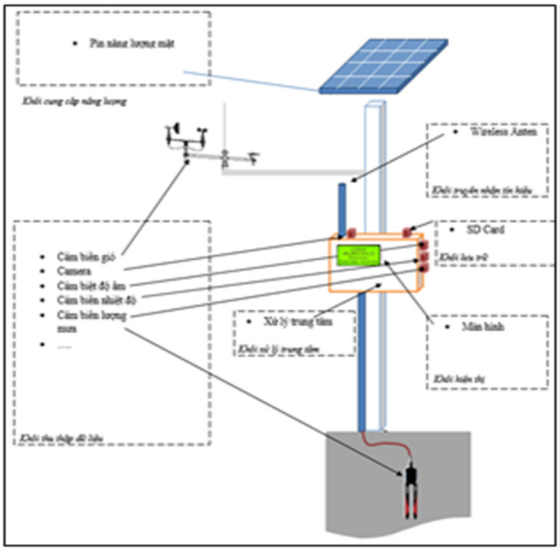
– Khối thu thập dữ liệu: bao gồm các cảm biến:

- Ánh sáng: thu thập cường độ ánh sáng.
- Nhiệt độ: thu thập nhiệt độ môi trường.
- Độ ẩm: thu thập các thông số độ ẩm của đất, độ ẩm không khí.
- Độ pH: đo độ pH của nước.

– Khối trung tâm xử lý: Đảm nhận việc tập hợp dữ liệu từ khối thu thập dữ liệu sau đó điều khiển khối giao tiếp để truyền về nút quản lý vùng, và ra lệnh cho khối hiện thị xuất ra màn hình.

– Khối giao tiếp: Sau khi khối trung tâm xử lý tập hợp dữ liệu và gửi cho khối giao tiếp, khối giao tiếp sẽ thực hiện công việc truyền dữ liệu về nút quản lý vùng.

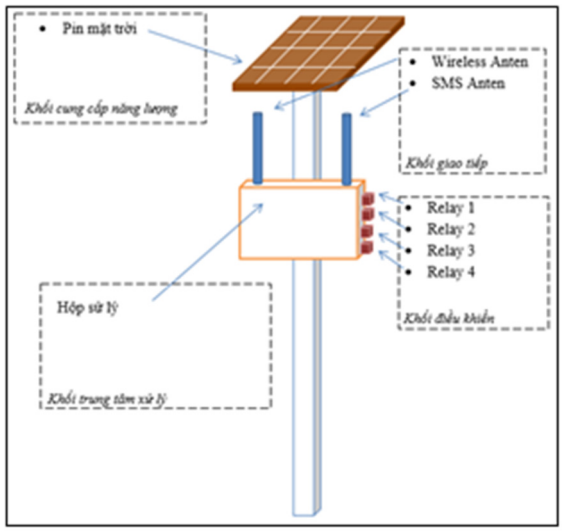
– Khối lưu trữ: Công việc duy nhất là lưu lại dữ liệu vào thẻ nhớ.



Hình 6: Cấu trúc nút cảm biến (DHLA-WSN)

b. Nút quản lý vùng DHLA-WMN

Nút trung gian DHLA-WMN là nơi nhận dữ liệu từ nút cảm biến rồi truyền về cho trung tâm điều hành và đồng thời nhận lệnh từ trung tâm điều hành truyền về cho nút cảm biến để thực thi.



Hình 7: Cấu trúc nút quản lý vùng (DHLA-WMN)

Hình 7 mô tả cấu trúc của một nút quản lý vùng. Một nút quản lý vùng 4 thành phần sau:

– Khối cung cấp năng lượng: khối này sẽ cấp nguồn điện 12V cho thiết bị hoạt động. Điện có thể lấy từ năng lượng mặt trời, pin dự trữ …

– Khối trung tâm xử lý: xử lý các thông tin nhận được từ khối giao tiếp và trả dữ liệu về nếu cần.

– Khối giao tiếp: đảm nhận công việc giao tiếp dữ các nút cảm biến và trung tâm điều hành.

– Khối điều khiển: thực hiện đóng, mở các thiết bị như máy bơm, máy phun sương, đèn, quạt gió…

3.3.2 Trung tâm điều hành

Trung tâm điều điều hành là nơi nhận và xử lý số liệu, tại trung tâm khi nhận dữ liệu sẽ được lưu và hiển thị trên PC, giúp các nhà sản xuất, nhà nghiên cứu có dữ liệu chính xác ngay tại thời điểm.

Với các dữ liệu được lưu trên PC và một số thuật toán phân tích số liệu đặc thù sẽ giúp cho nông dân, nhà sản xuất có một góc nhìn tổng thể về khu vực họ cần giám sát, hỗ trợ cho họ một số quyết định (hệ hỗ trợ quyết định)

Từ PC họ có thể điều khiển lập ra lịch cho hệ thống tưới tiêu hay hoạt động khi thỏa một số yêu

cầu mà họ đề ra. PC sẽ tự động gởi lệnh về hệ thống tưới tiêu.

Tại trung tâm có kết nối với Internet, GPRS, SMS hỗ trợ giám sát hoặc điều khiển từ xa. Hệ thống máy tính, thiết bị tại trung tâm DHLA_CENTER được xây dựng như sau:

Phần cứng

– Máy Server: chức năng quản lý các thiết bị ngoại vi.

– Modem SMS: gửi, nhận SMS qua các nhà mạng.

– Modem wireless: thiết bị thu phát dữ liệu không dây.

– Modem ADSL: thiết bị thu phát mạng băng thông rộng.

3.4 Phần mềm giám sát, điều kiển DHLA_SOFT

3.4.1 Tại các nút cảm biến (DHLA-WSN-SOFT)

Chương trình được xây dựng tại các nút cảm biến cứ sau 1 khoảng thời gian sẽ cảm biến môi trường, thông qua bộ lọc dữ liệu (data filter) và gởi về nút quản lý vùng. Thông tin có thể gởi đi với lịch trình định trước hoặc gởi đi khi nhận được yêu cầu từ nút quản lý vùng.

3.4.2 Nút quản lý vùng DHLA-WMN-SOFT

DHLA-WMN-SOFT có 2 chức năng chính: xử lý các thông tin nhận được từ các nút DHLA-WSN và truyền về trung tâm điều hành; tiếp nhận và xử lý các yêu cầu từ trung tâm điều hành với các nút DHLA-WSN.

3.4.3 Trung tâm điều hành

Phần mềm DHLA_SOFT được triển khai tại trung tâm điều hành hoặc tổng trung tâm điều hành. Phần mềm DHLA_SOFT gồm các chứng năng chính sau:

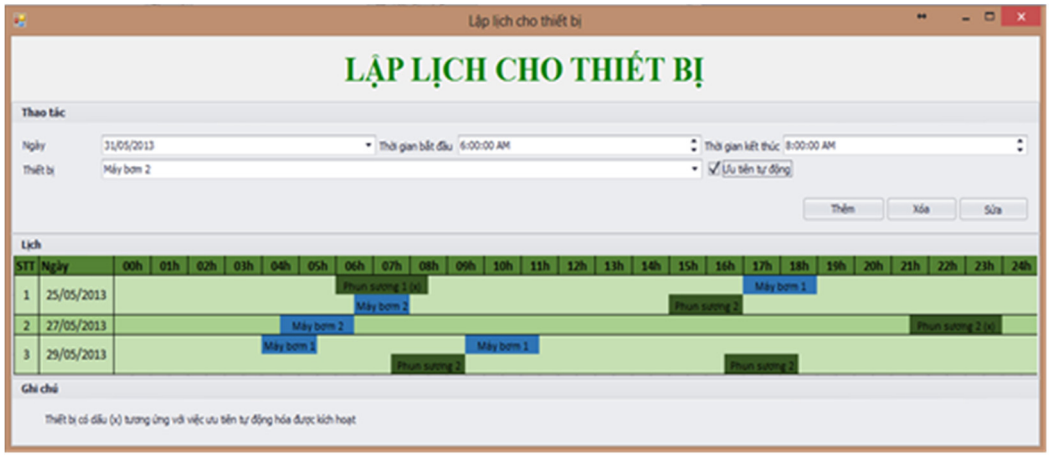
– Hiện thị trên bản đồ vùng giám sát (Hình 8).

– Hiển thị và lưu trữ thông dữ liệu thông số môi trường phục vụ việc theo dõi lâu dài cho người sử dụng (Hình 9).

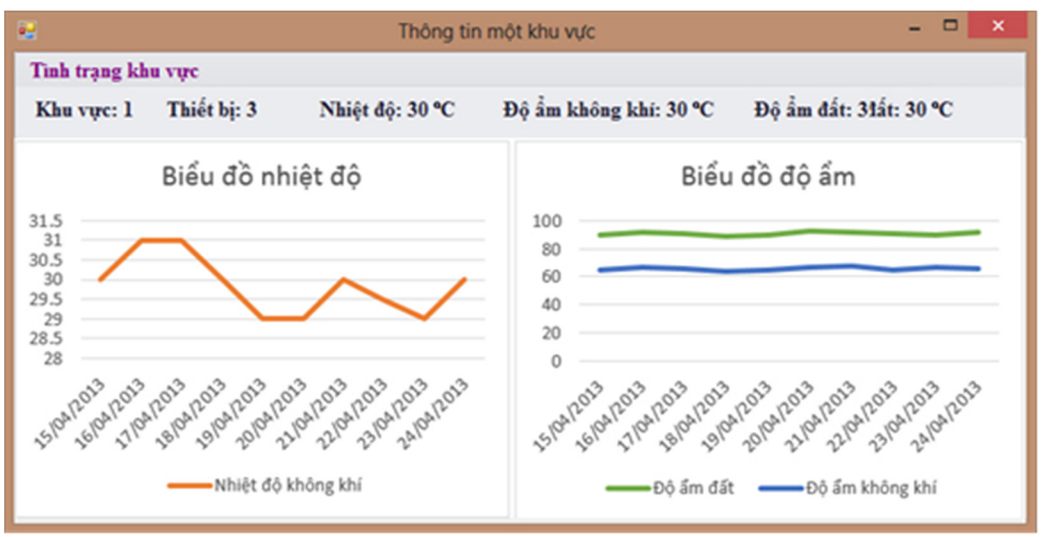
– Lập lịch và lập kế hoạch điều khiển tự động vùng giám sát quản lý (Hình 10). Các nhà nông dân sử dụng có thể lập kế hoạch cho hệ thống hoạt động theo một chu trình nào đó. Ví dụ hệ thống sẽ tưới tiêu từ 8 giờ đến 9 h hằng ngày từ thứ 2 đến thứ 7. Tuy nhiên việc lập lịch không thể linh hoạt theo sự thay đổi của môi trường, nên hệ thống phần mềm cũng cung cấp cho việc lập kế hoạch điều khiển tự động dựa vào những thông số mà nút quản lý vùng gửi về (Hình 11). Ví dụ khi nhiệt độ tăng cao hơn 30^0C thì hệ thống sẽ phun sương để làm mát môi trường và tăng độ ẩm, đến khi nhiệt độ hạ xuống 25^0C thì hệ thống sẽ tắt phun sương.

Hệ hỗ trợ ra quyết định, với dữ liệu mà hệ thống thu thập được theo thời gian, và dữ liệu về cây trồng thì phần mềm có thể giúp người dùng chăm sóc cây trồng tốt hơn.

Hình 8: Giao diện thông tin khu vực giám sát



Hình 9: Giao diện thông tin nhiệt độ và độ ẩm tại một nút cảm biến



Hình 10: Giao diện thiết lập kế hoạch tưới tiêu theo thời gian

Hình 11: Giao diện thiết lập điều khiển kiểm soát môi trường tự động

4 KẾT LUẬN

Bài viết này đã giới thiệu việc thiết kế, xây dựng và triển khai mạng cảm biến không dây phục vụ nông nghiệp chính xác tại Trường Đại học Kinh tế Công nghiệp Long An. Hệ thống đã được xây dựng, triển khai và chạy thử nghiệm tại trường bước đầu hoạt động tốt hứa hẹn đem lại nhiều lợi ích đáng kể. Đây là công nghệ mới giúp nông dân và doanh nghiệp dễ dàng tiếp cận và ứng dụng vào sản xuất nông nghiệp, tiết kiệm nguyên liệu đầu vào, nâng cao hiệu quả, năng suất và lợi nhuận trong hệ thống sản xuất nông nghiệp. Đặc biệt, với việc làm chủ công nghệ hoàn toàn từ việc thiết kế xây dựng phần cứng đến phần mềm cho các nút cảm biến hoạt động nên giá thành của sản phẩm không cao, ít phụ thuộc vào công nghệ nước ngoài và phù hợp với thị trường tại Việt Nam.

Mô hình này có thể được phát triển và ứng dụng rộng rãi cùng với các chuyên gia nông nghiệp để tìm ra bộ thông số tối ưu cho mỗi loại cây trồng phù
hợp cho mỗi địa phương, phát huy được tiềm năng và thế mạnh của sản xuất tại địa phương để tạo ra sản phẩm năng suất cao và giá trị được lợi nhuận trên
đầu tư của mình bằng cách tiết kiệm các chi phí có liên quan.

Mạng cảm biến không dây phục vụ cho nông nghiệp là một hướng đi mới cho nông nghiệp đồng bằng sông Cửu Long nói riêng và cả nước nói chung để hướng tới nông nghiệp bền vững. Nông nghiệp bền vững tìm cách để đảm bảo một nguồn cung cấp liên tục của thực phẩm trong giới hạn sinh thái, kinh tế và xã hội cần thiết để duy trì sản xuất trong dài hạn. Do đó, nông nghiệp chính xác bằng cách sử dụng mạng cảm biến không dây sẽ cho phép theo đuổi mục tiêu này.

LỜI CẢM TẠ

Chúng tôi xin chân thành cảm ơn Ban Lãnh đạo Trường Đại học Kinh tế Công nghiệp Long An đã tài trợ cho công trình nghiên cứu này. Chúng tôi cũng xin gởi lời cảm ơn đến Ông Samuel Raj, Tổng giám đốc công ty Seyerswide Pte Ltd, Singapore; TS. Wen Hu, Senior Research Scientist tại Viện Nghiên cứu Úc CSIRO, Brisbane, Australia; TS. Sarfraz Nawaz, Researcher tại Trường Đại học Cambridge, UK đã hỗ trợ và tư vấn kỹ thuật cho chúng tôi.

TÀI LIỆU THAM KHẢO

1. Xiong, S.-M.; Wang, L.-M.; Qu, X.-Q.; Zhan, Y.-Z. Application Research of WSN in Precise Agriculture Irrigation. In Proceedings of International Conference on Environmental Science and Information Application Technology, Wuhan, China, 4–5 July 2009; Volume 2, pp. 297-300.
2. Jennifer Yick, Biswanath Mukherjee, Dipak Ghosal, Wireless Sensor Network Survey, Computer Networks, August 2008, Volume 52, Issue 12, pp. 2292-2330.
3. Al-hamdi, Ali, Ahmed Monjurul Hasan, and Muhammad Akram. "WSN-based Support for Irrigation Efficiency Improvements in Arab Countries."
4. KSHITIJ SHINGHAL, ARTI NOOR, etc, Wireless Sensor Networks in agriculture: for potato farming.
5. de Lima, G.H.E.L.; e Silva, L.C.; Neto, P.F.R. WSN as a Tool for Supporting Agriculture in the Precision Irrigation. In Proceedings of International Conference on Networking and Services, Cancun, Mexico, 7–13 March 2010; pp. 137-142.
6. Kim, Y.; Evans, R.; Iversen, Remote Sensing and Control of an Irrigation System Using a Distributed Wireless Sensor Network. IEEE Trans. Instrum. Meas. 2008, 57, 1379-1387.
7. Aline Baggio, (2004), “Wireless sensor networks in precision agriculture”, The Netherlands IEEE Pervasive Computing, 3(1):38–45, Jan-Mar 2004.
8. B. S. Kang , S. Kim , F. Ren , B. P. Gila , C. R. Abernathy and S. J. Pearton, (2005), “Comparison of \MOS and Schottky W/Pt–GaN diodes for hydrogen detection,” *Sens. Actuators,Chem.,* vol.104, pp.232, 2005.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

TƯ VẤN HỌC TẬP NHÓM BẬC ĐẠI HỌC TRÊN CƠ SỞ TIẾP CẬN LỚP TƯƠNG ĐỒNG LỚN NHẤT TRONG HỆ THỐNG THÔNG TIN ĐA TRỊ

Phan Tấn Tài[1], Trần Minh Tân[1], Phạm Ngọc Quyền[1], Nguyễn Thị Thu Thảo[1] và Huỳnh Xuân Hiệp[1,2]

[1] Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ
[2] Nhóm nghiên cứu liên ngành DREAM (Decision-support Research for Environmental Applications and Models), Trường Đại học Cần Thơ

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Academic advising undergraduate groups based on the approach of maximal tolerance class of set-valued information system

Từ khóa:
Hệ thống thông tin đa trị, Lớp tương đồng, Chia sẻ thông tin, Kết quả học tập, Kết quả rèn luyện

Keywords:
Set-valued information system, Tolerance class, Information-sharing, Academic performance, Personal training

ABSTRACT

This paper proposes a new approach for academic advising, especially for undergraduate groups, based on the approach of maximal tolerance class in the set-valued information system to build groups of students with similar learning outcomes and training results. Based on these results, the academic advisors can organize and propose solutions for each group of similar students in an effective manner, and help the students with poor academic performance to find the solutions to overcome their weakness. At the same time, the academic advisors can also identify the similar groups of good learning outcome students so that the advisors can build groups for studying and doing scientific research to provide opportunities for students getting better in studying.

TÓM TẮT

Bài viết này giới thiệu một tiếp cận mới về tư vấn học tập nhóm bậc đại học dựa trên khái niệm lớp tương đồng lớn nhất (maximal tolerance class) của hệ thống thông tin đa trị (set-valued information system) để xây dựng các nhóm sinh viên có tương đồng về kết quả học tập, kết quả rèn luyện. Từ đó, cố vấn học tập có thể tổ chức tư vấn và đề ra những giải pháp cho mỗi nhóm sinh viên tương đồng một cách hiệu quả, giúp các sinh viên có kết quả học tập, rèn luyện yếu kém có hướng khắc phục tốt hơn. Đồng thời qua đó, cố vấn học tập cũng xác định được các nhóm sinh viên tương đồng về kết quả học tập và rèn luyện tốt, từ đó có thể xây dựng các nhóm học tập, nghiên cứu khoa học nhằm tạo cơ hội cho sinh viên không ngừng tiến bộ.

1 GIỚI THIỆU

Xây dựng mô hình chia sẻ thông tin phục vụ cố vấn học tập (CVHT) trong công tác sinh viên là một vấn đề cấp thiết hiện nay, nhất là tại các trường đã và đang chuyển sang mô hình đào tạo theo hệ thống tín chỉ. Một mô hình chia sẻ thông tin như vậy cần được nghiên cứu một cách cơ bản để hình thành các kênh chia sẻ thông tin phục vụ CVHT một cách hiệu quả trên nền công nghệ thông tin. Mỗi kênh thông tin sẽ gắn kết với một chủ thể cung cấp thông tin phục vụ công tác CVHT nhằm đảm bảo CVHT có đủ thông tin về sinh viên, đồng thời qua đó có thể tổ chức tư vấn sinh viên một cách tốt nhất.

Trong bài viết này, chúng tôi đặt trọng tâm vào việc xây dựng các nhóm sinh viên tương đồng về kết quả học tập dựa trên cơ sở tiếp cận lớp tương đồng lớn nhất của hệ thống thông tin đa trị nhằm giúp CVHTcó thể tư vấn và đưa ra các giải pháp hỗ trợ kịp thời để sinh viên có những định hướng đúng đắn trong học tập và rèn luyện.

Bài báo được giới thiệu gồm 6 phần: phần thứ nhất giới thiệu tổng quan, phần thứ hai trình bày về hệ thống thông tin đa trị, phần ba giới thiệu các dạng thông tin phục vụ CVHT, phần thứ tư trình bày vấn đề về tư vấn học tập nhóm bậc đại học, phần thứ năm trình bày thực nghiệm và sau cùng là phần kết luận.

2 HỆ THỐNG THÔNG TIN ĐA TRI

2.1 Hệ thống thông tin đa trị

Một hệ thống thông tin [1][3] được định nghĩa gồm 4 bộ sau: <*U,Q,V,f*>. Trong đó, U là một tập hữu hạn không rỗng gồm N đối tượng $\{x_1, x_2, ..., x_N\}$, Q là một tập hữu hạn không rỗng gồm n thuộc tính $\{q_1, q_2, ..., q_n\}$, $V = \bigcup_{q \in Q} \{V\}_q$ với V_q là miền giá trị của thuộc tính q, $f: UxQ \rightarrow V$ là hàm thông tin, nghĩa là $f(x,q) \in V_q$.

Ví dụ 1: xem xét một tập dữ liệu đơn giản cho trong Bảng 1, các tập U, Q, V, f được xác định như sau:

$U=\{x_1, x_2, x_3, x_4\}$,

$Q=\{q_1, q_2, q_3, q_4\}$,

$V_{q_1} = \{0,1,2,2\}$ (các giá trị trên cột q_1 của Bảng 1),

$V_{q_2} = \{1,2,1,1\}$ (các giá trị trên thuộc tính q_2 của Bảng 1),…và $f(x_1, q_1) = 0$, $f(x_1, q_2) = 1$, …

Bảng 1: Hệ thống thông tin S = <*U,Q,V,f*>

U	q_1	q_2	q_3	q_4
x_1	0	1	0	1
x_2	1	2	0	0
x_3	2	1	1	0
x_4	2	1	2	0

Nếu ứng với mỗi đối tượng $x_{i\,(i=1,2,...,N)}$, mỗi thuộc tính tương ứng trong hệ thống thông tin <*U,Q,V,f*> có một giá trị duy nhất (a unique attribute value) thì <*U,Q,V,f*> được gọi là hệ thống thông tin đơn trị (a single-valued information system) và ngược lại <*U,Q,V,f*> được gọi là hệ thống thông tin đa trị (set-valued (multi-valued) information system).

Ví dụ 2: xem xét một tập dữ liệu đơn giản cho hệ thống thông tin <*U,Q,V,f*> trong Bảng 2, các giá trị của hàm thông tin được biểu diễn như sau: $f(x_1, q_1) = \{0,1\}$, $f(x_2, q_1) = \{0,2\}$,… Nghĩa là, mỗi thuộc tính trong hệ thống thông tin đa trị có thể có một hay nhiều giá trị ứng với một đối tượng x_i, $i = 1,2,...N$. Trong trường hợp này <*U,Q,V,f*> được gọi là một hệ thống thông tin đa trị.

Bảng 2: Hệ thống thông tin đa trị gồm 4 đối tượng {x_1, x_2, x_3, x_4} và 4 thuộc tính {q_1, q_2, q_3, q_4}

U	q_1	q_2	q_3	q_4
x_1	{0, 1}	1	0	1
x_2	{0, 2}	{0, 1}	2	1
x_3	{0, 2}	{1, 2}	1	0
x_4	2	1	2	0

2.2 Lớp tương đồng

Lớp tương *đồng (tolerance class)* [3]: Trong một hệ thống thông tin đa trị <*U,Q,V,f*> , ta định nghĩa $T_B(x) = \{y \mid y \in U, \forall b \in B: b(x) \cap b(y) \neq \emptyset\}$ là một lớp tương đồng của $x \in U$ dựa trên tập thuộc tính điều kiện $B \subseteq Q$

Theo Bảng 3, nếu gọi $B=Q$

$Q = \{q_1, q_2, q_3, q_4, q_5\}$, khi đó ta có:

$T_B(x_1) = T_B(x_7) = \{x_1, x_7, x_9\}$,

$T_B(x_2) = T_B(x_8) = \{x_2, x_3, x_8\}$,

$T_B(x_3) = \{x_2, x_3, x_4, x_8\}$,

$T_B(x_4) = \{x_3, x_4\}$,

$T_B(x_5) = T_B(x_6) = \{x_5, x_6\}$,

$T_B(x_9) = \{x_1, x_7, x_9, x_{10}\}$.

Bảng 3: Một hệ thống thông tin quyết định đa trị với 10 đối tượng $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$

U	q_1	q_2	q_3	q_4	q_5
x_1	{1}	{0,1}	{0}	{1,2}	{2}
x_2	{0,1}	{2}	{1,2}	{0}	{0}
x_3	{0}	{1,2}	{1}	{0,1}	{0}
x_4	{0}	{1}	{1}	{1}	{0, 2}
x_5	{2}	{1}	{0,1}	{0}	{1}
x_6	{0,2}	{1}	{0,1}	{0}	{1}
x_7	{1}	{0,2}	{0,1}	{1}	{2}
x_8	{0}	{2}	{1}	{0}	{0,1}
x_9	{1}	{0,1}	{0,2}	{1}	{2}
x_{10}	{1}	{1}	{2}	{0,1}	{2}

Khi phân lớp các đối tượng của tập *U* dựa trên khái niệm lớp tương đồng (tolerance class) có thể dẫn đến hai trường hợp cần xem xét sau: (1) các đối tượng trong $T_Q(x)$ có thể không cùng có các giá trị thuộc tính; (2) với $x \neq y; x, y \in U$, có thể tồn tại quan hệ bao hàm nhau giữa $T_Q(x)$ và $T_Q(y)$. Xem lại ví dụ 3: (1) Trong $T_Q(x_3)$, x_2 và x_4 có hai thuộc tính q_2 và q_5 không cùng có giá trị giống nhau giữa x_2 và x_4; (2) Ta có các quan hệ tương đồng bao hàm: $T_Q(x_2) \in T_Q(x_3)$, $T_Q(x_{10}) \in T_Q(x_9)$.

Để giải quyết vấn đề trên, chúng ta phân lớp các đối tượng thành các lớp tương đồng lớn nhất (maximal tolerance class).

Lớp tương ***đồng lớn nhất*** là lớp tương đồng mà trong đó mỗi đối tượng đều tương đồng với các đối tượng khác.

Bảng 4: Các lớp tương đồng lớn nhất cùng các giá trị tương đồng

Lớp tương đồng (maximal tolerance class)	Giá trị tương đồng
$\{x_1, x_7, x_9\}$	(1, 0, 0, 1, 2)
$\{x_2, x_3, x_8\}$	(0, 2, 1, 0, 0)
$\{x_3, x_4\}$	(0, 1, 1, 1, 0)
$\{x_5, x_6\}$	(2, 1, {0, 1}, 0, 1)
$\{x_9, x_{10}\}$	(1, 1, 2, 1, 2)

3 THÔNG TIN PHỤC VỤ CVHT

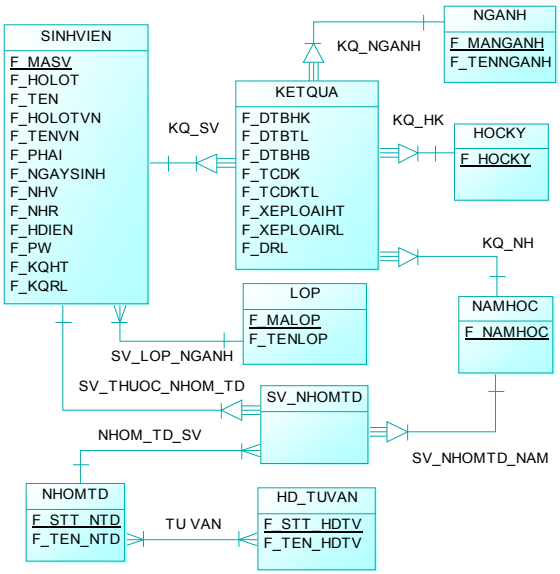
Thông tin phục vụ CVHT được xác định theo các dạng sau [7][10]: thông tin chung về sinh viên; thông tin khen thưởng, kỷ luật; thông tin vi phạm qui định, qui chế học vụ; kết quả học tập, kết quả rèn luyện (Điểm rèn luyện); thông tin về điểm cộng, điểm trừ của các tổ chức Đoàn thể; thông tin đoàn viên ưu tú và kết quả phân loại đoàn viên; thông tin sinh viên sinh hoạt công dân (sinh hoạt đầu năm, lao động, học tập chính trị,...); thông tin cảnh báo học vụ, buộc thôi học, thông tin nợ học phí, học trễ tiến độ.

4 TƯ VẤN HỌC TẬP NHÓM BẬC ĐẠI HỌC

Với cách tiếp cận mới về lớp tương đồng lớn nhất của hệ thống thông tin đa trị, kết quả có được từ sự phân lớp tương đồng lớn nhất dựa trên kết quả học tập sẽ cho ra nhiều lớp tương đồng lớn nhất mà trong đó có thể cùng một sinh viên xuất hiện ở hai lớp tương đồng lớn nhất khác nhau. Ví dụ với sinh viên A có kết quả học tập hai học kỳ là {"kém", "yếu"} thì sinh viên A sẽ có mặt ở cả hai lớp tương đồng "kém" và "yếu" (nếu chỉ phân lớp tương đồng dựa trên một thuộc tính kết quả học tập). Tuy nhiên, vấn đề đặt ra cho việc tư vấn học tập nhóm bậc đại học là chúng tôi cần xác định: (1) các nhóm sinh viên tương đồng lớn nhất về kết quả học tập, mà trong đó mỗi sinh viên chỉ có ở một nhóm tương đồng lớn nhất và (2) các hoạt động tư vấn cần thiết cho mỗi nhóm sinh viên tương đồng. Do đó, việc phân lớp tương đồng lớn nhất dựa trên kết quả học tập được thực hiện như sau: từ kết quả học tập của hai học kỳ trong năm học, những sinh viên chỉ cần có kết quả học lực một học kỳ đạt loại "kém" thì những sinh viên này được xếp vào nhóm sinh viên tương đồng "kém"(những sinh viên này không được xếp vào các nhóm tương đồng tiếp theo). Tương tự, những sinh viên chỉ cần có kết quả học lực một học kỳ đạt loại "yếu" thì những sinh viên đó được xếp vào nhóm sinh viên tương đồng "yếu" và phân nhóm tương tự cho các nhóm tương đồng "trung bình", "khá" và "giỏi-xuất sắc".

4.1 Mô hình dữ liệu mức quan niệm phục vụ tư vấn học tập nhóm

Mô hình dữ liệu mức quan niệm liên quan đến kết quả học tập của sinh viên được thể hiện như Hình 1 [4], thực thể kết quả thể hiện các kết quả tổng hợp về học tập và rèn luyện của sinh viên. Dữ liệu của các kết quả này được tổng hợp từ hệ thống quản lý đào tạo (QLĐT) của Trường Đại học Cần Thơ.



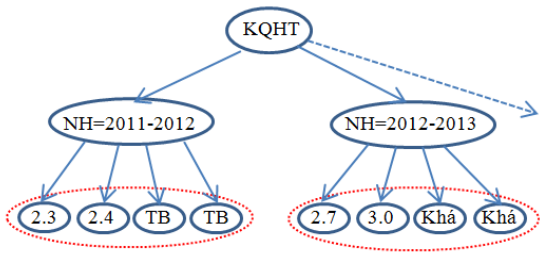
Hình 1: Mô hình dữ liệu mức quan niệm cho việc phân nhóm tương đồng và phục vụ hoạt động tư vấn sinh viên

Theo Hình 1, một điểm mới trên mô hình này là chúng tôi đưa vào thực thể sinh viên (SINHVIEN) hai thuộc tính kết quả học tập (KQHT) và kết quả rèn luyện (KQRL). *Hai thuộc tính KQHT và KQRL là các thuộc tính đa trị, cả hai thuộc tính đều có kiểu dữ liệu XML.* Với cách thiết kế này, kết quả học tập và kết quả rèn luyện của sinh viên sẽ được lưu trữ dưới dạng XML.Cũng theo Hình 1: mỗi sinh viên ở một năm học thuộc một nhóm tương đồng (kém, yếu, trung bình, khá hay giỏi-xuất sắc), mỗi nhóm tương đồng sẽ có những hoạt động tư vấn và mỗi hoạt động tư vấn cũng có thể được thực hiện ở một hay nhiều nhóm tương đồng.

4.2 Cấu trúc dữ liệu XML cho các thuộc tính đa trị

Cấu trúc dữ liệu dạng XML[2][9] cho thuộc tính KQHT có dạng như sau:

```
<NAMHOCNAMHOC="2011-2012 " />
<NAMHOCNAMHOC="2011-2012 ">
<TBHK>
<TBHK>2.3</TBHK>
</TBHK>
<TBHK>
<TBHK>2.4</TBHK>
</TBHK>
<XEPLOAIHT>
<XEPLOAIHT>TB</XEPLOAIHT>
</XEPLOAIHT>
<XEPLOAIHT>
<XEPLOAIHT>TB</XEPLOAIHT>
</XEPLOAIHT>
</NAMHOC>
```

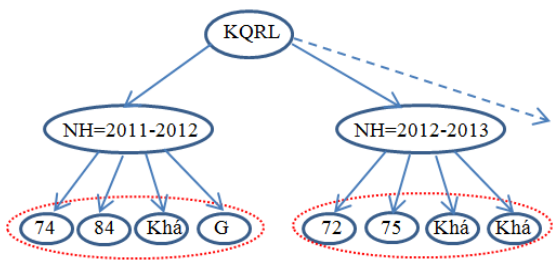


Hình 2: Cây cấu trúc dữ liệu dạng XML cho thuộc tính KQHT

Theo Hình 2, kết quả học tập của mỗi sinh viên bao gồm điểm trung bình và xếp loại học lực hai học kỳ của mỗi năm học sẽ được bổ sung vào thuộc tính KQHT dưới dạng một nút chứa bốn nút lá trên cây cấu trúc dữ liệu XML. Theo đó, kết quả học tập năm 2011-2012 của sinh viên được hiểu như sau: điểm trung bình học kỳ 1 là 2.3 và học kỳ 2 là 2.4, xếp loại học lực học kỳ 1 là "TB" và học kỳ 2 là "TB".

Cấu trúc dữ liệu dạng XML [2][9]cho thuộc tính KQRL có dạng như sau:

```
<NAMHOCNAMHOC="2011-2012 " />
<NAMHOCNAMHOC="2011-2012 ">
<DRL>
<DRL>74</DRL>
</DRL>
<DRL>
<DRL>84</DRL>
</DRL>
<XEPLOAIRL>
<XEPLOAIRL>Khá</XEPLOAIRL>
</XEPLOAIRL>
<XEPLOAIRL>
<XEPLOAIRL>G</XEPLOAIRL>
</XEPLOAIRL>
</NAMHOC>
```



Hình 3: Cây cấu trúc dữ liệu dạng XML cho thuộc tính KQRL

Theo Hình 3, kết quả rèn luyện của mỗi sinh viên bao gồm điểm rèn luyện và xếp loại rèn luyện hai học kỳ của mỗi năm học sẽ được bổ sung vào thuộc tính KQRL dưới dạng một nút chứa bốn nút lá trên cây cấu trúc dữ liệu XML. Theo đó, kết quả rèn luyện năm 2011-2012 của sinh viên được hiểu

là: điểm rèn luyện học kỳ 1 là 74 và học kỳ 2 là 84, xếp loại học lực học kỳ 1 là "Khá" và học kỳ 2 là "G".

4.3 Phân nhóm sinh viên dựa trên tiếp cận hệ thống thông tin đa trị

Với tiếp cận của hệ thống thông tin đa trị, ta gọi U = {Tập các sinh viên}, $Q = \{q_1, q_2\} = \{KQHT, KQRL\}$ ={Tập thuộc tính}, V là miền giá trị của các thuộc tính KQHT và KQRL (KQHT: Kết quả học tập, KQRL: Kết quả rèn luyện). Với thuộc tính KQHT và thuộc tính KQRL là các thuộc tính đa trị. f là hàm thông tin, phản ánh kết quả học tập và rèn luyện của các sinh viên. Khi đó một hệ thống thông tin đa trị sẽ là một bộ gồm tập ký hiệu như sau: <*U,Q,V,f*>.

Ví dụ: xét hệ thống thông tin đa trị với $U = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}\}$ (tương ứng với 10 sinh viên) và tập thuộc tính $Q = \{q_1, q_2\} = \{KQHT, KQRL\}$

Bảng 5: Hệ thống thông tin đa trị <*U,Q,V,f*> với dữ liệu là kết quả học tập và rèn luyện trong một năm học

U	q_1 (KQHT)	q_2(KQRL)
s_1	{3, 2}	{3, 3}
s_2	{2, 2}	{2, 1}
s_3	{2, 2}	{2, 2}
s_4	{2, 3}	{3, 3}
s_5	{4, 4}	{4, 4}
s_6	{1, 1}	{1, 2}
s_7	{2, 1}	{2, 2}
s_8	{4, 3}	{4, 4}
s_9	{1, 1}	{1, 1}
s_{10}	{2, 2}	{3, 2}

Qui ước xếp loại: 4 = Giỏi, 3 = Khá, 2 = Trung bình và 1 = Yếu

Khi đó, hàm thông tin $f(s_i, q_j)$ được biểu diễn như sau: $f(s_1, q_1) = \{3,2\}$, $f(s_1, q_2) = \{3,3\}$, $f(s_2, q_1) = \{2,2\}$, $f(s_2, q_2) = \{2,1\}$,...

Đặt $B_1 = \{q_1\}, B_2 = \{q_2\}$, $B = \{q_1, q_2\}$ và thực hiện phân lớp tập U thành các nhóm sinh viên tương đồng lần lượt trên B_1, B_2 và B.

4.3.1 *Phân nhóm sinh viên tương đồng chỉ dựa trên kết quả học tập* $(B)_1$

– Nhóm sinh viên tương đồng kết quả học tập xếp loại "yếu" $(T_{\downarrow}(B_{\downarrow}1, Y))$:

$$T_{B_1,Y} = (T_{\downarrow}(B_{\downarrow}1, Ymax)) = \{s_6, s_7, s_9\}$$

$T_{B_1,Y}$ *là nhóm sinh viên có kết quả học tập yếu ở một học kỳ hay cả hai học kỳ.*

Bảng 6. Hệ thống thông tin đa trị <*U,Q,V,f*> với việc phân nhóm sinh viên tương đồng về kết quả học tập trong một năm học

U	q_1 (KQHT)
s_1	{3, 2}
s_2	{2, 2}
s_3	{2, 2}
s_4	{2, 3}
s_5	{4, 4}
s_6	{1, 1}
s_7	{2, 1}
s_8	{4, 3}
s_9	{1, 1}
s_{10}	{2, 2}

– Nhóm sinh viên tương đồng kết quả học tập xếp loại "trung bình" ở một học kỳ hay cả hai học kỳ nhưng không có kết quả học tập xếp loại "yếu" $(T_{\downarrow}(B_{\downarrow}1, TB))$:

$$T_{B_1,TB} = (T_{\downarrow}(B_{\downarrow}1, TBmax)) - T_{\downarrow}(B_{\downarrow}1, Y)$$

Với $(T_{\downarrow}(B_{\downarrow}1, TBmax)) = \{s_1, s_2, s_3, s_4, s_7, s_{10}\}$ là nhóm sinh viên tương đồng lớn nhất về kết quả học tập xếp loại "trung bình". Vậy,

$$T_{B_1,TB} = \{s_1, s_2, s_3, s_4, s_7, s_{10}\} - \{s_6, s_7, s_9\} = \{s_1, s_2, s_3, s_4, s_{10}\}$$

– Nhóm sinh viên tương đồng kết quả học tập xếp loại "khá" ở một học kỳ hay cả hai học kỳ trong năm học nhưng không có kết quả học tập "yếu" và "trung bình" $(T_{\downarrow}(B_{\downarrow}1, K))$:

$$T_{B_1,K} = (T_{\downarrow}(B_{\downarrow}1, Kmax)) - (T_{B_1,TB} + T_{\downarrow}(B_{\downarrow}1, Y))$$
$$= \{s_1, s_4, s_8\} - (\{s_{\downarrow}1, s_{\downarrow}2, s_{\downarrow}3, s_{\downarrow}4, s_{\downarrow}10\} + \{s_{\downarrow}6, s_{\downarrow}7, s_{\downarrow}9\}) = \{s_8\}$$

– Nhóm sinh viên tương đồng kết quả học tập xếp loại "giỏi" ở cả hai học kỳ $(T_{\downarrow}(B_{\downarrow}1, G))$:

$$T_{B_1,G} = (T_{\downarrow}(B_{\downarrow}1, Gmax)) - (T_{B_1,K} + T_{B_1,TB} + T_{B_1,Y})$$
$$= \{s_5, s_8\} - (\{s_8\} + \{s_1, s_2, s_3, s_4, s_{10}\} + \{s_6, s_7, s_9\}) = \{s_5\}$$

Bảng 7: Kết quả phân nhóm sinh viên tương đồng lớn nhất về kết quả học tập trong năm

Nhóm tương đồng lớn nhất	Giá trị tương đồng về KQHT
$\{s_6, s_7, s_9\}$	1
$\{s_1, s_2, s_3, s_4, s_{10}\}$	2
$\{s_8\}$	3
$\{s_5\}$	4

4.3.2 Phân nhóm sinh viên tương đồng dựa trên kết quả học tập $((B)_1)$ và kết quả rèn luyện $((B)_2)$

Theo Bảng 5, nếu phân nhóm dựa trên hai kết quả học tập và rèn luyện của một năm học ta có:

– Nhóm sinh viên tương đồng về kết quả học tập và rèn luyện xếp loại "yếu" $(T_1(B,Y))$:

$$T_{B,Y} = (T_1(B,Ymax)) = \{s_6, s_9\}$$

$T_{B,Y}$ *là nhóm sinh viên có kết quả học tập và rèn luyện yếu ở một học kỳ hay cả hai học kỳ.*

– Nhóm sinh viên tương đồng có kết quả học tập và rèn luyện"trung bình" ở một học kỳ hay cả hai học kỳ nhưng không có kết quả học tập và rèn luyện yếu $(T_1(B,TB))$:

$$T_{B,TB} = (T_1(B,TBmax)) - T_1(B,Y)$$

Với $(T_1(B,TBmax)) = \{s_2, s_3, s_7, s_{10}\}$ là nhóm sinh viên tương đồng lớn nhất về kết quả học tập và rèn luyện có xếp loại "trung bình". Vậy,

$$T_{B,TB} = \{s_2, s_3, s_7, s_{10}\} - \{s_6, s_9\}$$
$$= \{s_2, s_3, s_7, s_{10}\}$$

– Nhóm sinh viên tương đồng có kết quả học tập và rèn luyện xếp loại "khá" ở một học kỳ hay cả hai học kỳ nhưng không có kết quả học tập và rèn luyện "yếu" và "trung bình" $(T_1(B,K))$:

$$T_{B,K} = (T_1(B,Kmax)) - (T_{B,TB} + T_1(B,Y))$$
$$= \{s_1, s_4\} - (\{s_1 2, s_1 3, s_1 7, s_1 10\} + \{s_1 6, s_1 9\})$$
$$= \{s_1, s_4\}$$

– Nhóm sinh viên tương đồng có kết quả học tập và rèn luyện xếp loại "giỏi" ở cả hai học kỳ ($T_1(B,G)$) :

$$T_{B,G} = (T_1(B,Gmax)) - (T_{B,K} + T_{B,TB} + T_{B,Y})$$
$$= \{s_5, s_8\} - (\{s_1, s_4\} + \{s_2, s_3, s_7, s_{10}\} + \{s_6, s_9\})$$
$$= \{s_5, s_8\}$$

Bảng 8: Kết quả phân nhóm sinh viên tương đồng về kết quả học tập và rèn luyện trong năm

Nhóm tương đồng lớn nhất	Giá trị tương đồng về KQHT và KQRL
$\{s_6, s_9\}$	(1, 1)
$\{s_2, s_3, s_7, s_{10}\}$	(2, 2)
$\{s_1, s_4\}$	(3, 3)
$\{s_5, s_8\}$	(4, 4)

4.4 Các hoạt động tư vấn

Việc phân nhóm sinh viên dựa trên lớp tương đồng lớn nhất về kết quả học tập đã chỉ ra các nhóm sinh viên tương đồng về học lực. Từ đây, CVHT có thể thực hiện các hoạt động tư vấn phù hợp với từng nhóm sinh viên tương đồng nhằm góp phần nâng cao chất lượng đào tạo của nhà trường.

Bảng 9: Một số hoạt động tư vấn học tập nhóm bậc đại học

STT	Tên hoạt động tư vấn học tập
1	(HĐ1): Tổ chức họp nhóm sinh viên tương đồng để thu thập thông tin, xác định nguyên nhân và tư vấn giải pháp khắc phục.
2	(HĐ2):Rà soát và điều chỉnh kế hoạch.
3	(HĐ3): Xây dựng kế hoạch họp nhóm sinh viên tương đồng theo định kỳ.
4	(HĐ4): Tổ chức họp nhóm sinh viên tương đồng theo kế hoạch.
5	(HĐ5): Xây dựng nhóm sinh viên học tập
6	(HĐ6): Tổ chức tư vấn về tâm tư tình cảm, thái độ học tập.
7	(HĐ7): Tổ chức tư vấn, chia sẻ kinh nghiệm trong học tập.
8	(HĐ8): Tư vấn chương trình đào tạo, đề cương học phần.
9	(HĐ9): Tư vấn tổ chức các nhóm nghiên cứu khoa học trong sinh viên.
10	(HĐ10): Tổ chức tư vấn một số kỹ năng cơ bản về giao tiếp giữa sinh viên với sinh viên, giữa sinh viên với thầy cô.
11	(HĐ11): Tổ chức tư vấn về các hoạt động đoàn và phong trào lớp.
12	(HĐ12): Tổ chức tư vấn về việc tham gia các phong trào có ích cho cộng đồng.

Theo Bảng 9, (HĐ1): giúp CVHT và sinh viên tìm ra được những nguyên nhân hạn chế và đưa ra các giải pháp khắc phục. (HĐ2): giúp sinh viên rà

soát và điều chỉnh kế hoạch cho phù hợp với năng lực hiện tại của sinh viên. (HĐ3): nhằm xác định kế hoạch để CVHT theo dõi và quản lý chặt chẽ đối với các nhóm sinh viên tương đồng (đặc biệt là nhóm học tập yếu, kém). (HĐ4): nhằm mục tiêu tư vấn cách giải quyết vấn đề và chia sẻ những khó khăn sinh viên gặp phải trong quá trình học.(HĐ5): nhằm tạo điều kiện để các sinh viên giúp đỡ lẫn nhau trong học tập. (HĐ6): giúp các sinh viên tự tin và có ý chí vươn lên trong học tập. (HĐ7): giúp sinh viên nâng cao kết quả học tập. (HĐ8): là một hoạt động hết sức cần thiết nhằm giúp sinh viên hiểu rõ về chương trình đào tạo và đề cương học phần. (HĐ9): nhằm đẩy mạnh nghiên cứu khoa học trong sinh viên, qua đó giúp sinh viên phát triển về mặt học thuật và nâng cao khả năng tư duy sáng tạo hơn nữa. (HĐ10): giúp sinh viên tự tin hơn trong các hoạt động giao tiếp. (HĐ11): giúp sinh viên có nhận thức đúng đắn về các phong trào đoàn, phong trào lớp. Qua đó, góp phần tạo điều kiện để sinh viên yêu thích tham gia các hoạt động xã hội một cách tích cực trong quá trình rèn luyện nhân cách con người. (HĐ12): giúp sinh viên rèn luyện ý thức tự nguyện và có trách nhiệm với cộng đồng.

5 THỰC NGHIỆM

Nội dung này sẽ giới thiệu một số vấn đề tư vấn cho các nhóm sinh viên tương đồng về kết quả học tập. Một số so sánh, phân tích giữa các cách xây dựng nhóm tương đồng cũng được trình bày.

5.1 Dữ liệu thực nghiệm

5.1.1 Dữ liệu ban đầu

Dữ liệu ban đầu là dữ liệu tổng hợp về kết quả học tập và rèn luyện học kỳ 1 và học kỳ 2, năm 2011-2012 của lớp DIXX95A1 (Hệ thống thông tin K_XX). Các dữ liệu này được kết xuất sang hai tập tin HK1_1112.xls và HK2_1112.xls) [11].

Bảng tổng hợp điểm học kỳ 1, năm học 2011-2012 gồm 14 cột lần lượt có tên sau: Stt, Mã SV, Họ và tên, Ng/Sinh, MH01, MH02, MH03, MH04, TCHK, TBHK, D.RL, TCTL, TBCH, TBHB. Bảng gồm có 56 dòng tương ứng với tổng số 56 sinh viên.

Bảng tổng hợp điểm học kỳ 2, năm học 2011-2012 gồm 18 cột lần lượt có tên sau: Stt, Mã SV,Họ và tên, Ng/Sinh, MH01, MH02, MH03, MH04, MH05, MH06, MH07, MH08, TCHK, TBHK,D.RL, TCTL, TBCH, TBHB. Bảng gồm có 56 dòng tương ứng với tổng số 56 sinh viên.

5.1.2 Tinh chỉnh dữ liệu phục vụ phân nhóm sinh viên

Các cột dữ liệu của bảng tổng hợp điểm học kỳ như: Mã SV, Họ và tên, TCHK, TBHK, D.RL, TCTL, TBCH, TBHB là các cột dữ liệu được chọn đưa vào thực nghiệm, *đồng thời mã số và tên sinh viên đã được thay đổi để đảm bảo tính bảo mật dữ liệu riêng tư của sinh viên*.

Các bước tinh chỉnh dữ liệu được thực hiện như sau: Bước 1, Tạo tập tin DIXX95A1.xls có dòng đầu tên DIXX95A1, dòng thứ hai là tiêu đề cột (gồm các cột: Mã SV, HK, NH, MANGANH, TCHK,TBHK, XLHL, D.RL, XLRL, TCTL,TBCH, TBHB), dòng thứ ba trở đi lưu dữ liệu kết quả học tập và rèn luyện tương ứng (như Hình 4); Bước 2, chép các dữ liệu quan tâm từ các cột: Mã SV, TCHK, TBHK, D.RL, TCTL, TBCH, TBHB của hai tập tin tên HK1_1112.xls và HK2_1112.xls sang tập tin DIXX95A1.xls, đồng thời điền dữ liệu (học kỳ 1, 2 năm học 2011-2012, mã ngành 95) lấp đầy cho các cột HK, NH, MANGANH; Bước 3, thực hiện đưa dữ liệu trên tập tin DIXX95A1.xls vào bảng KQHT của cơ sở dữ liệu tích hợp phục vụ CVHT (bằng chức năng *Import* phần mềm SIS2A).

	A	B	C	D	E	F	G	H	I	J	K	L
1	DIXX95A1											
2	**Mã SV**	**HK**	**NH**	**MANG**	**TCHK**	**TBHK**	**XLHL**	**D.RL**	**XLRL**	**TCTL**	**TBCH**	**TBHB**
3	1111201	1	2011-2012	95	15	2.33	TB	74	Khá	16	2.33	
4	1111201	2	2011-2012	95	17	2.44	TB	84	G	33	2.30	2.44
5	1111202	1	2011-2012	95	15	2.00	TB	60	TB	15	2.00	2.00
6	…											

Hình 4: Cấu trúc bảng tập tin DIXX95A1.xls

5.2 Công cụ sử dụng SIS2A

Theo Hình 5, SIS2A là phần mềm hệ thống thông tin đa trị chia sẻ thông tin phục vụ cố vấn học tập (Set-valued Information Systems for Academic Advisor (SIS2A) tool) [8].

Hình 5: Giao diện hệ thống quản lý và chia sẻ thông tin phục vụ cố vấn học tập [8]

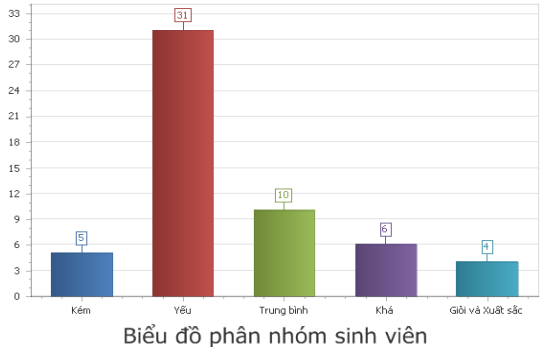
Phần mềm SIS2A gồm 3 phân hệ chính: quản lý sinh viên (dành cho phòng CTSV), quản lý đoàn viên (dành cho Đoàn khoa), quản lý chia sẻ thông tin sinh viên (dành cho CVHT).

Ngoài ra, còn có nhóm chức năng phân nhóm sinh viên nhằm hỗ trợ CVHT tổ chức tư vấn cho sinh viên. Theo Hình 6, nhóm chức năng phân nhóm sinh viên gồm có các chức năng chính như: Import dữ liệu phân nhóm (đây là chức năng mở rộng), chức năng phân nhóm (classification), chức năng kết xuất báo cáo.

Với chức năng phân nhóm theo Hình 8, CVHT có thể thực hiện các tùy chọn khác nhau để thực hiện phân nhóm sinh viên tương đồng như chọn lớp chuyên ngành, năm học, tiêu chí phân nhóm (phân nhóm theo kết quả học tập hay phân nhóm theo kết quả rèn luyện) và giá trị tương đồng (kém, yếu, trung bình, khá, giỏi, xuất sắc hoặc chọn tất cả).

Hình 6: Chức năng phân nhóm sinh viên tương đồng (*Classification*) về kết quả học tập và rèn luyện [8]

Theo Hình 7, kết quả thống kê cho thấy có 5 sinh viên có học lực "kém", 31 sinh viên có học lực "yếu", 10 sinh viên có học lực "trung bình", 6 sinh viên có học lực "khá" và 4 sinh viên có học lực "giỏi-xuất sắc".



Hình 7: Biểu đồ thống kê số lượng sinh viên của các nhóm tương đồng về kết quả học tập [**8**]

5.3 Tư vấn nhóm sinh viên thuộc lớp tương đồng "kém"

5.3.1 Phân tích kết quả phân nhóm sinh viên có học lực "kém"

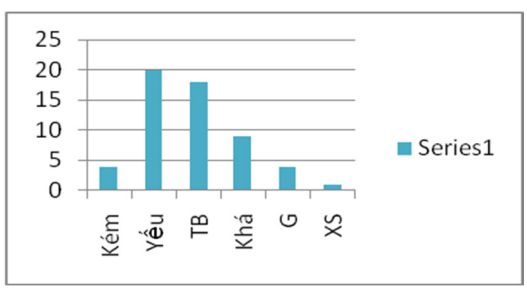
Theo Hình 8, việc phân nhóm sinh viên "kém" dựa trên cơ sở tiếp cận lớp tương đồng lớn nhất cho kết quả gồm có 5 sinh viên.

+ NHOM SINH VIEN : Kém

STT	MÃ SV	HỌ TÊN	THÔNG TIN HỌC TẬP				
			HK	TBHK	XLHT	DRL	XLRL
1	1111208	Nguyễn Đăng K	1	1.13	Y	50	TB
			2	0.37	Kém	43	Y
			HK	TBHK	XLHT	DRL	XLRL
2	1111214	Triệu Hoàng N	1	1.13	Y	56	TB
			2	0.81	Kém	59	TB
			HK	TBHK	XLHT	DRL	XLRL
3	1111220	Nguyễn Thanh S	1	0.27	Kém	50	TB
			2	1.03	Y	50	TB
			HK	TBHK	XLHT	DRL	XLRL
4	1111227	Nguyễn Lê Trí T	1	1.00	Y	55	TB
			2	0.74	Kém	48	Y
			HK	TBHK	XLHT	DRL	XLRL
5	1117847	Nguyễn Phước S	1	1.13	Y	65	TB
			2	0.89	Kém	49	Y

Hình 8: Kết quả phân nhóm sinh viên "kém" trên cơ sở tiếp cận lớp tương đồng lớn nhất [8]

Hình 9 cho kết quả thống kê xếp loại học tập cả năm của lớp DIXX95A1 dựa trên điểm trung bình của hai học kỳ. Theo đó, số sinh viên có học lực "kém" là 4, số sinh viên có học lực "yếu" là 20, số sinh viên có học lực "trung bình" là 18, số sinh viên có học lực "khá" là 9, số sinh viên có học lực "giỏi" là 4 và số sinh viên có học lực "xuất sắc" là 1.



Hình 9: Biểu đồ thống kê kết quả xếp loại học tập cả năm của lớp DIXX95A1 dựa trên điểm trung bình của hai học kỳ

Theo Bảng 10, kết quả xếp loại và thống kê xếp loại học tập cả năm của lớp DIXX95A1 được thực hiện dựa trên điểm trung bình của hai học kỳ cho thấy nhóm sinh viên có học lực "kém" chỉ có 4 sinh viên, không có sinh viên Nguyễn Phước S (Mã SV: 1117847).

Trong khi đó, với cách phân nhóm dựa trên cơ sở tiếp cận lớp tương đồng lớn nhất của hệ thống thông tin đa trị th sinh viên Nguyễn Phước S thuộc nhóm sinh viên "kém". Bởi vì, theo kết quả kết xuất chi tiết trên Hình 10 cho thấy, kết quả học tập của sinh viên Nguyễn Phước S ở hai học kỳ lần lượt có xếp lại "yếu" và "kém" và theo cách xác định lớp tương đồng về kết quả học tập (đã trình bày ở mục 4.3.1) thì sinh viên nào đó chỉ cần có kết quả học tập xếp loại "kém" ở một trong hai học

kỳ (hay cả hai học kỳ) thì sinh viên đó được đưa vào nhóm tương đồng có kết quả học tập "kém".

Bảng 10: Kết quả xếp loại và phân nhóm sinh viên "kém" dựa trên công cụ Microsoft Excel

STT	Mã SV	TB HK1	TB HK2	TB CN	XL CN
1	1111208	0.37	1.13	0.75	Kém
2	1111214	0.81	1.13	0.97	Kém
3	1111220	1.03	0.27	0.65	Kém
4	1111227	0.74	1.00	0.87	Kém

Với cách xác định nhóm sinh viên tương đồng có kết quả học tập "kém" như vậy thì sinh viên Nguyễn Phước S được xem như một sinh viên thuộc diện "kém"mà CVHT phải quan tâm tư vấn đặc biệt. Trong khi đó nếu thống kê dựa vào điểm trung bình cả năm thì sinh viên Nguyễn Phước S không thuộc diện sinh viên có học lực "kém" để CVHT quan tâm tư vấn đặc biệt. Điều này có thể dẫn đến sự thiếu sót trong việc quan tâm tư vấn đặc biệt cho sinh viên S. Như vậy dẫn đến khả năng sinh viên này có thể tiếp tục học kém và phải bị cảnh báo học vụ hay thậm chí bị buộc thôi học. Do đó, việc phân nhóm sinh viên tương đồng dựa trên cở sở tiếp cận lớp tương đồng lớn nhất của hệ thống thông tin đa trị như trên sẽ giúp CVHT xác định các nhóm sinh viên học tập "kém" một cách hợp lý và hiệu quả hơn.

5.3.2 Thực hiện tư vấn nhóm sinh viên học tập "kém"

Khi đã xác định được nhóm sinh viên tương đồng về học lực "kém", CVHT có thể thực hiện ngay một số hoạt động cần thiết như HĐ1, HĐ2, HĐ3, HĐ4, HĐ5, HĐ6, HĐ7. Các bước tiến hành: Bước thứ nhất thực hiện HĐ1 để thu thập thông tin và tìm ra nguyên nhân học kém của sinh viên; Bước thứ hai thực hiện HĐ2, CVHT cùng sinh viên có học lực "kém" rà soát và điều chỉnh kế hoạch tiếp theo cho phù hợp với năng lực của mỗi sinh viên; Bước thứ ba thực hiện HĐ3 và HĐ4 để xây dựng kế hoạch họp nhóm, nghe sinh viên báo cáo tiến độ học tập, tư vấn cách giải quyết vấn đề và chia sẻ những khó khăn sinh viên gặp phải trong quá trình học; Bước thứ tư thực hiện HĐ5, CVHT xây dựng các nhóm sinh viên học tập để giúp đỡ lẫn nhau trong quá trình học; Bước thứ năm thực hiện HĐ6 và HĐ7, CVHT gặp riêng các sinh viên này để tư vấn về tâm tư tình cảm, thái độ học tập và nhân cách để giúp các em tự tin và có ý chí vươn lên trong học tập và rèn luyện.

5.4 Tư vấn nhóm sinh viên thuộc lớp tương đồng "Yếu"

5.4.1 Phân tích kết quả phân nhóm sinh viên có học lực "yếu"

Theo Hình 7, kết quả thống kê số lượng sinh viên của các nhóm tương đồng về kết quả học tập dựa trên cơ sở tiếp cận lớp tương đồng lớn nhất thì lớp DIXX95A1 có 31 sinh viên thuộc nhóm tương đồng có học lực "yếu". Trong khi đó (theo Hình 9), kết quả thống kê dựa trên kết quả học tập trung bình cả năm cho thấy chỉ có 20 sinh viên có học lực "yếu". Hai kết quả thống kê cho kết quả khác nhau là vì: (1) với cách phân nhóm dựa vào điểm trung bình cả năm, ta không quan tâm đến điểm trung bình của mỗi học kỳ; (2) với cách phân nhóm dựa trên tiếp cận lớp tương đồng lớn nhất thì những sinh viên nào có kết quả học tập ở một trong hai học kỳ xếp loại "yếu"(và không thuộc nhóm sinh viên có học lực "kém") sẽ được đưa vào nhóm tương đồng học tập "yếu".

Như vậy, với cách phân nhóm dựa trên tiếp cận lớp tương đồng lớn nhất thì nhóm sinh viên tương đồng "yếu" xác định được có tính đầy đủ, tránh việc thiếu quan tâm đến những sinh viên có tiềm ẩn kết quả học tập "yếu" mà không được CVHT quan tâm đúng mức.

5.4.2 Thực hiện tư vấn nhóm sinh viên học tập "yếu"

Việc thực hiện tư vấn nhóm sinh viên học tập "yếu" tương tự như thực hiện tư vấn đối với sinh viên học tập "kém". Bởi vì, nhóm sinh viên học tập "yếu" cũng có tiềm ẩn cao tiếp tục có kết quả học tập "yếu" hay "kém".

5.5 Tư vấn nhóm sinh viên thuộc lớp tương đồng "giỏi-xuất sắc"

5.5.1 Phân tích kết quả phân nhóm sinh viên có học lực "giỏi-xuất sắc"

Theo Bảng 11, kết quả thống kê xếp loại học tập cả năm của lớp DIXX95A1 dựa theo kết quả điểm trung bình cộng của hai học kỳ cho thấy số sinh viên có học lực "giỏi-xuất sắc" là 5.

Bảng 11: Kết quả phân nhóm sinh viên có học lực "giỏi" dựa theo điểm trung bình cả năm

STT	Mã SV	TB HK1	TB HK2	TB CN	XL CN
1	1111211	3.27	3.23	3.25	G
2	**1111215**	**3.66**	**2.73**	**3.20**	**G**
3	1117733	3.63	3.43	3.53	G
4	1117843	3.85	3.47	3.66	XS
5	1117850	3.69	3.53	3.61	G

Theo Hình 10, việc phân nhóm sinh viên "Giỏi-Xuất sắc" dựa trên cơ sở tiếp cận lớp tương đồng lớn nhất cho kết quả chỉ có 4 sinh viên thuộc nhóm có học lực "giỏi-xuất sắc", không có sinh viên có mã số11111215. Bởi vì theo cách phân nhóm dựa trên khái niệm lớp tương đồng lớn nhất thì sinh viên có mã số 11111215 được xem như thuộc nhóm sinh viên có tương đồng về kết quả học tập "khá". Với quan niệm đó, nhóm sinh viên tương đồng có học lực "giỏi-xuất sắc" là đáng tin cậy, bởi vì cả hai học kỳ điều đạt học lực "giỏi" trở lên. Mặt khác nếu xem xét chi tiết hơn thì sinh viên có mã số 11111215 đang có kết quả học tập đi xuống bất thường (HK1: 3.66, xuất sắc; HK2: 2.73, khá gần cận trung bình). Do đó, CVHT cần quan tâm nhiều đối với sinh viên này.

+ NHÓM SINH VIÊN : Giỏi và Xuất sắc

STT	MÃ SV	HỌ TÊN	THÔNG TIN HỌC TẬP				
			HK	TBHK	XLHT	DRL	XLRL
1	1111211	Lâm Quang Kha L	1	3.23	G	91	XS
			2	3.27	G	93	XS
			HK	TBHK	XLHT	DRL	XLRL
2	1117733	Phạm Quang Huy B	1	3.43	G	90	XS
			2	3.63	XS	90	XS
			HK	TBHK	XLHT	DRL	XLRL
3	1117843	Bùi Cẩm N	1	3.47	G	81	G
			2	3.85	XS	90	XS
			HK	TBHK	XLHT	DRL	XLRL
4	1117850	Đỗ Lê Nhật T	1	3.53	G	71	Khá
			2	3.69	XS	80	G

Ngày in: 10/6/2013 7:53:11 PM

Hình 10: Kết quả phân nhóm sinh viên "giỏi-xuất sắc" trên cơ sở tiếp cận lớp tương đồng lớn nhất [8]

5.5.2 Thực hiện tư vấn nhóm sinh viên học tập "giỏi-xuất sắc"

Đối với nhóm sinh viên tương đồng về kết học tập "giỏi-xuất sắc" theo Hình 12, CVHT có thể thực hiện một số hoạt động tư vấn cần thiết như: (HĐ8): Tư vấn chương trình đào tạo, đề cương học phần nhằm giúp sinh viên lựa chọn các học phần tốt nhất có thể cho nghề nghiệp và phát huy tối đa khả năng của sinh viên (ví dụ như khuyến khích sinh viên tham gia làm đề tài luận văn tốt nghiệp thay vì học môn thay thế). (HĐ9): Tư vấn tổ chức các nghiên cứu khoa học trong sinh viên nhằm mục tiêu đẩy mạnh nghiên cứu khoa học trong sinh viên, qua đó giúp sinh viên phát triển về mặt học thuật và nâng cao khả năng tư duy sáng tạo hơn nữa, hoạt động này phù hợp với nhóm sinh viên tương đồng giỏi, xuất sắc. (HĐ10): Tổ chức tư vấn một số kỹ năng cơ bản về giao tiếp như thuyết trình, trao đổi, thảo luận, phản biện. (HĐ11): Tổ chức tư vấn về các hoạt động Đoàn giúp sinh viên có nhận thức đúng đắn về các phong trào đoàn, phong trào lớp và tham gia mạnh mẽ các công tác Đoàn, công tác xã hội.(HĐ12): Tổ chức tư vấn về việc tham gia các phong trào có ích cho cộng đồng nhằm giúp sinh viên rèn luyện ý thức tự nguyên, trách nhiệm và tình thương cộng đồng, từ đó góp phần xây dựng cộng đồng tốt đẹp.

6 KẾT LUẬN

Chúng tôi đã tiếp cận vấn đề dựa trên khái niệm lớp tương đồng lớn nhất (maximal tolerance class) của hệ thống thông tin đa trị (set-valued information system)[1][3] để xây dựng chức năng phân nhóm sinh viên có tương đồng về kết quả học tập và rèn luyện. Từ đó, giúp CVHT có thể tư vấn và đề ra những giải pháp cho mỗi nhóm sinh viên tương đồng một cách hiệu quả. Đặc biệt, CVHT có thể xác định được nhóm sinh viên học tập "kém" và "yếu" từ đó thực hiện một số hoạt động tuvấn phù hợp giúp các sinh viên này có hướng khắc phục tốt hơn. Đối với nhóm sinh viên học tập "giỏi-xuất sắc", CVHTcó thể thực hiện một số hoạt động tư vấn, tạo điều kiện để các sinh viên này tham gia nghiên cứu khoa học và các hoạt động khác nhằm tạo cơ hội cho sinh viên không ngừng tiến bộ về mọi mặt. Lợi ích khi giải quyết các vấn đề dựa trên tiếp cận của hệ thống thông tin đa trị là: (1) Có thể vận dụng để xây dựng các thuộc tính đa trị với việc lưu trữ dữ liệu dạng XML và có thể truy xuất dữ liệu một cách hiệu quả; (2) Kết quả xếp loại học tập và rèn luyện mỗi năm học của sinh viên có thể lưu trữ ngay trên bảng sinh viên nhờ các thuộc tính đa trị.

Phần mền SIS2A cho phép phân nhóm sinh viên tương đồng và kết xuất chi tiết kết quả học tập và rèn luyện của sinh viên theo mỗi nhóm tương đồng nhưng chưa cho phép xác định nhóm sinh viên tương đồng có kết quả học tập giảm. Do đó, hướng phát triển tiếp của phần mềm SIS2A là cải tiến chức năng phân nhóm sinh viên tương đồng cho phép xác định thêm nhóm sinh viên tương đồng có kết quả học tập giảm dựa trên thứ tự kết quả đạt được của hai học kỳ.

LỜI CẢM TẠ

Xin chân thành cảm ơn sự góp ý của Phòng Công tác Sinh viên và Đoàn Khoa CNTT&TT – Trường Đại học Cần Thơ. Nội dung nghiên cứu này được triển khai trong đề tài cấp cơ sở "Nghiên cứu xây dựng mô hình quản lý và chia sẻ thông tin phục vụ công tác cố vấn học tập".

TÀI LIỆU THAM KHẢO

1. Cios, K. J., Pedrycz, W., Swiniarski, R.W.: Rough Sets: Data Mining: Methods for Knowledge Discovery, pp 27-45.

Kluwer Academic Publishers, Boston / Dordrecht / London (1998).

2. Deitel, H.M., Deitel, P. J., Nieto, T. R., Internet & Word Wide Web How to Program,pp 633-680. Prentice Hall (Second Edition).
3. Guan, Y.-Y., Wang, H.-K.: Set-valued Information Systems, International Journal of Information Sciences 176, 2507-2525 (2006)
4. Hoffer, J., George, J. F., Valacich, J. S.:Modern System Analysis & Design. Prentice Hall (2002)
5. Nguyễn Anh Tuấn, Quy định về công tác học vụ (ban hành kèm theo quyết định số 1210/QĐ-ĐHCT, ngày 06/8/2008 của Hiệu trưởng trường ĐHCT), Trường Đại học Cần Thơ, 2008.
6. Nguyễn Anh Tuấn, Quy định công tác cố vấn học tập (ban hành kèm theo Quyết định số 2067/QĐ-ĐHCT, ngày 04 tháng 12 năm 2007 của Hiệu trưởng Trường Đại học Cần Thơ), Trường Đại học Cần Thơ, 2007.
7. Nguyễn Anh Tuấn, Sổ tay giảng viên, Trường Đại học Cần Thơ, 2010.
8. Phan Tấn Tài, "Nghiên cứu xây dựng mô hình quản lý và chia sẻ thông tin phục vụ công tác cố vấn học tập" , mã số: T2013-03, đề tài nghiên cứu cấp cơ sở, Trường Đại học Cần Thơ, 2013.
9. Raul Chong, Ian Hakes, Rav Ahuja. DB2 Express-C, pp187-210. IBM (2008).
10. Trần Thị Thanh Hiền, Một số vấn đề trong công tác quản lý sinh viên (Biên bản họp thống nhất một số vấn đề trong công tác quản lý sinh viên, số 2284/BB-ĐHCT, ngày 18 tháng 12 năm 2012), Trường Đại học Cần Thơ, 2012.
11. Hệ thống quản lý đào tạo, (https://qldiem.ctu.edu.vn/htql/canbo/qldiem/codes/index.php?mID=60501), Trường Đại học Cần Thơ.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

GIẢI THUẬT ĐƠN GIẢN ĐỂ PHÁT HIỆN LÀN ĐƯỜNG VÀ ĐIỀU KHIỂN LÁI CHO ÔTÔ TỰ HÀNH

Trương Quốc Bảo[1]
[1] *Khoa Công nghệ, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Simple lane detection and steering control algorithm for autonomous guided vehicle

Từ khóa:
Ôtô tự hành (AVG), Phát hiện làn đường, Vecto-làn đường, Non-uniform Bspline (NUBS), Độ cong của làn đường

Keywords:
Automated Guided Vehicle (AGV), Lane detection, Vector-lane-concept, Non-uniform B-spline (NUBS), Road lane curvature

ABSTRACT

Automated Guided Vehicle (AGV) is an intelligent machine that has enough intelligence to determine its motion status according to the environment conditions. TypicalAGV has the ability to operate and move purposely without human intervention. In this paper, we present a simple method based on computer vision techniques which help the vehicle to move inside the lane boundaries. The vision image will be processed to detect the lane boundaries using vector-lane-concept and estimate the road lane curvature. Using this information, we calculate the steering angle which is used to steer the vehicle to move inside the lane boundaries without using any other control technique. Several demonstrations were carried out, using diverse images, to show the result of lane-boundary detection and vehicle navigation. In conclusion, the proposed algorithm can be used to control vehicle to move inside the lane boundaries without human intervention.

TÓM TẮT

Ôtô tự hành (AVG) là một dạng máy thông minh có đủ khả năng hiểu biết để tự xác định trạng thái chuyển động của mình dựa trên những điều kiện môi trường. Một dạng tiêu biểu của ôtô hay robot tự hành là khả năng vận hành và chuyển động không cần sự điều khiển của con người. Trong bài báo này, chúng tôi trình bày một giải thuật xử lý ảnh đơn giản giúp ôtô tự di chuyển trong giới hạn làn đường của mình. Giải thuật sẽ tiến hành phát hiện là đường sử dụng khái niệm vecto-làn đường và ước lượng độ cong của làn đường. Dựa trên các thông tin nhận được chúng tôi tính góc lái và tiến hành xây dựng giải thuật lái xe giúp ôtô tự hành di chuyển trong giới hạn là đường của mình mà không cần sử dụng thêm giải thuật điều khiển nào. Nhiều dạng khác nhau của ảnh làn đường đã được sử dụng để minh họa cho tính hiệu quả của giải thuật được đề nghị. Các kết quả thực nghiệm đã cho thấy giải thuật có thể được áp dụng để điều khiển ôtô tự hành di chuyển trong giới hạn là đường của mình mà không cần tác động của con người.

1 GIỚI THIỆU

Ôtô tự hành được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau trong thực tế và đóng vai trò quan trọng trong môi trường giao thông có nguy cơ tai nạn cao như đường đồi dốc, trơn trợt hay các hệ thống giao thông thông minh. Điều khiển chuyển động tự động thông minh của xe có thể giúp hạn chế tai nạn cũng như ùn tắt giao thông.

Ôtô hay robot tự hành là lĩnh vực nghiên cứu thú vị và nhiều thách thức trong lĩnh vực trí tuệ nhân tạo và điều khiển học. Về nguyên tắc, một hệ

thống AVG đơn giản sẽ bao gồm hai thành phần chính: (1) Bộ phận tiền xử lý và (2) Bộ phận điều khiển.

Bộ tiền xử lý sử dụng hệ thống cảm biến (sensors), radar, GPS hoặc hệ thị giác máy tính với các camera đi kèm để thu nhận thông tin từ môi trường như giới hạn, hướng, độ rộng, độ cong, độ phẳng của làn đường cũng như sự xuất hiện của các vật cản,… Hầu hết các nghiên cứu gần đây đều sử dụng camera để nhận dạng làn đường. Nhiều mô hình và giải pháp khác nhau đã được đề nghị nhằm giải quyết vấn đề này. Đa phần các hệ thống đều nhận dạng làn đường dựa trên các giả thuyết định trước về mô hình toán học biểu diễn cho làn đường. Mô hình thường được sử dụng rộng rãi là đường cong bậc hai [1] hoặc Bsnack [2]. Gần đây cũng có một số nghiên cứu mới [3-4] nhưng vẫn chưa đạt được kết quả tối ưu. Vì vậy, bài toán phát hiện và nhận dạng làn đường vẫn là đề tài cần tiếp tục nghiên cứu và phát triển.

Bộ phận điều khiển sẽ điều khiển ôtô hay robot tự hành chuyển động dọc theo giới hạn làn đường của mình và có thể tránh các vật cản trên đường nếu cần. Có nhiều phương pháp có thể được áp dụng để xây dựng hệ thống điều khiển. Tiếp cận phổ biến nhất là hệ điều khiển mờ dựa trên fuzzy logic [5-6] hoặc hệ thống lai kết hợp giữa mạng neural và logic mờ [7]. Các hệ thống này thường là cồng kềnh tốn nhiều bộ nhớ và thời gian xử lý.

Trong nghiên cứu này, chúng tôi đề xuất một giải thuật đơn giản cùng mô hình tính toán thích nghi dựa trên các kỹ thuật xử lý ảnh nhằm giúp ôtô tự hành có thể di chuyển và điều khiển lái để chỉ trong giới hạn làn đường của mình tùy theo độ cong của làn đường. Mục đích chính là tìm một phương pháp hiệu quả, đơn giản nhưng phát hiện chính xác làn đường, đồng thời giảm thời gian và chi phí tính toán cho hệ thống điều khiển ôtô tự hành di chuyển chính xác theo làn đường quy định. Đầu tiên, chúng tôi sử dụng khái niệm vectơ làn đường dựa trên lý thuyết Non-uniform Bspline (NUBS) để xây dựng các đường giới hạn cho làn đường trái và phải. Tiếp theo, chúng tôi ước lượng độ cong của các làn đường trái và phải một cách riêng biệt. Cuối cùng, dựa trên thông tin về giới hạn và độ cong của làn đường, chúng tôi tính toán góc lái để sử dụng cho giải thuật điều khiển lái. Kết quả của quá trình xử lý ảnh sẽ cung cấp thông tin cần thiết và ước lượng độ cong của làn đường nhằm giúp ôtô tự hành di chuyển trong giới hạn làn đường quy định và xác định được hướng di chuyển chính xác cho ôtô từ vị trí hiện tại trên làn đường.

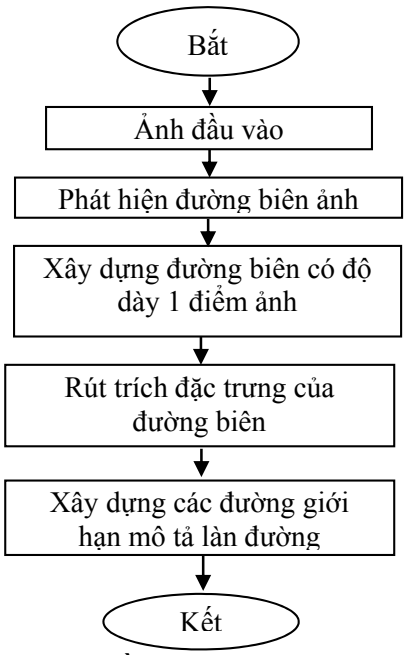
Các phần tiếp theo của bài báo được tổ chức như sau: Phần 2 chúng tôi đề xuất khái niệm vectơ là đường và phương pháp phát hiện và xây dựng giới hạn là đường dựa trên khái niệm này kết hợp với phép nội suy NUBS để đạt được bản đồ làn đường từ ảnh đầu vào. Giải thuật ước lượng độ cong của làn đường sẽ được trình bày trong mục 3. Tiếp đến, phương pháp điều khiển góc lái để cho phép ôtô tự hành di chuyển đúng làn đường quy định được mô tả chi tiết trong mục 4. Một số kết quả mô phỏng minh họa cho hệ thống được miêu tả và phân tích trong mục 5. Cuối cùng, Kết luận và định hướng nghiên cứu tương lai được nêu tóm tắt trong phần 6.

2 THUẬT TOÁN PHÁT HIỆN LÀN ĐƯỜNG

Các thuật toán xử lý ảnh cho phát hiện làn đường có thể được phân chia thành ba nhóm chính: (1) dựa trên phân vùng (region-based), (2) dựa trên phân tích đặc trưng (feature-based) và (3) dựa trên mô hình toán học hóa làn đường (model-based). Hầu hết các nghiên cứu đều đi theo hướng tiếp thứ ba này, sử dụng mô hình toán học để mô hình hóa làn đường [1-4]. Mục đích của phương pháp tiếp cận này là cố gắng làm cho phù hợp giữa ảnh làn đường được quan sát với mô hình giả thuyết được xây dựng. Kỹ thuật này dựa trên giả thuyết rằng làn đường là sự kết hợp của các phần có dạng đường thẳng, đường cong prabol hoặc đường cong dạng Bsnack.

Giải thuật cho phương pháp phát hiện làn đường của chúng tôi được trình bày như trong sơ đồ Hình 1. Đây là một hướng tiếp cận dựa trên sự mở rộng của mô hình hóa làn đường.

Đầu tiên, các điểm ảnh thuộc đường biên ảnh giới hạn làn đường sẽ được phát hiện nhờ thuật toán phát hiện đường biên Canny. Thuật toán rút xương ảnh [8] sẽ được áp dụng vào ảnh phát hiện cạnh để được ảnh đường biên có độ dày 1 điểm ảnh. Tiếp đến, ảnh đường biên sau khi rút xương sẽ được dùng để rút trích các điểm điều khiển (control points). Phép nội suy NBUS được sử dụng để xây dựng đường cong đại diện cho làn đường trái và phải một cách riêng biệt. Trong nghiên cứu của chúng tôi, các làn đường trái và phải được phát hiện và xây dựng một cách tách biệt. Điều này có nghĩa là các đường biên trái và phải được xem xét và xử lý riêng lẻ, không giống như nhiều nghiên cứu khác xem xét 2 làn đường trái và phải là đồng nhất với nhau.



Hình 1: Lưu đồ giải thuật phát hiện làn đường

2.1 Mô hình hóa làn đường bằng đường cong B-Spline

Đường cong NUBS ký hiệu là C [9] có thể được biển diễn bằng tập hợp $(n+1)$ điểm điều khiển $\{P_0, P_1, P_2, \cdots, P_n\}$. Điểm $P_i(x_i, y_i)$ thuộc đường cong C được tính theo công thức (1):

$$C(u) = \sum_{i=0}^{n} B_{i,m}(u) P_i \quad (1)$$

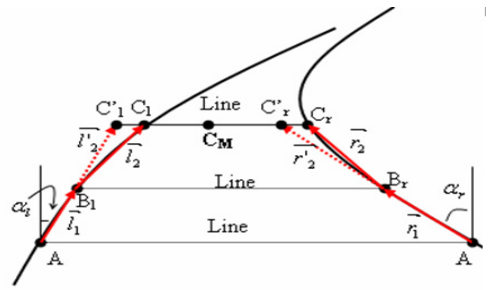
Trong đó, $t_{\min} \le u \le t_{\max}$, $2 \le m \le n+1$ và $B_{i,m}(u)$ được gọi là đường cong B-Spline và được định nghĩa như công thức (2):

$$B_{i,m}(u) = \begin{cases} \begin{cases} 1, t_i \le u \le t_{i+1} \\ 0, u \in [t_i, t_{i+1}] \end{cases}, & m=1 \\ \dfrac{u-t_i}{t_{i+m-1}-t_i} B_{i,m-1}(u) + \dfrac{t_{i+m}-u}{t_{i+m}-t_{i+1}} B_{i+1,m-1}(u), & m>1 \end{cases} \quad (2)$$

Xét công thức (2) ta thấy, để tính được giá trị của $B_{i,m}(u)$ chúng ta cần xác định được tập hợp các vectơ nút (knot vectos) $T = \{T_0, T_1, T_2, \cdots, T_{i+m}\}$ và các điểm nút (knot points) t_j, $j = \overline{0..i+m}$. Khi khoảng cách giữa các điểm nút là bằng nhau ta gọi đường cong là đường cong B-Spline đồng nhất (uniform), ngược lại là đường cong B-Spline không đồng nhất (Non-uniform). Càng nhiều điểm điều khiển và điểm nút được xác định thì đường cong B-Spline được biểu diễn bằng các đa thức bậc càng cao và sẽ mô tả đường cong càng chính xác.

2.2 Rút trích đặc trưng đường biên

Chúng tôi đề xuất mô hình làn đường như mô tả ở Hình 2 mà chúng tôi gọi là khái niệm **vectơ-làn đường (vecto-lane concept).**



Hình 2: Mô hình vectơ-làn đường cho bài toán phát hiện làn đường

Đầu tiên, chúng tôi xét 2 đường quét ngang trong ảnh bản đồ làn đường đạt được từ ảnh phát hiện đường biên và định nghĩa các điểm điều khiển là A_l, B_l cho làn đường bên trái; A_r, B_r cho làn đường bên phải. Xây dựng các vectơ làn đường $\vec{l}_1 = \overrightarrow{A_l B_l}$ cho làn đường bên trái, $\vec{r}_1 = \overrightarrow{A_r B_r}$ cho làn đường bên phải. Tiếp theo, chúng tôi ước lượng góc biểu diễn độ cong của đường biên biểu diễn làn đường sử dụng công thức (3):

$$\alpha_l = arctg\left(\frac{x_{A_l} - x_{B_l}}{y_{A_l} - y_{B_l}}\right); \alpha_r = arctg\left(\frac{x_{A_r} - x_{B_r}}{y_{A_r} - y_{Br}}\right) \quad (3)$$

Sử dụng góc biểu diễn độ cong của làn đường được định nghĩa trong công thức (3) ở trên, chúng tôi phát triển giải thuật để tìm kiếm các điểm điều khiển trên đường cong B-Spline. Nội dung chi tiết của giải thuật này được mô tả như sau:

Giải thuật tìm các điểm điều khiển

Bước 1: Thiết lập 2 đường quét ngang tại vùng trống ở vị trí cuối ảnh, tìm 2 điểm điều khiển đầu tiên cho mỗi đường quét, điểm điều khiển trái và điểm điều khiển phải. Vị trí bắt đầu quét để tìm các điểm điều khiển được chọn là vị trí trung điểm của mỗi đường quét.

Bước 2: Xây dựng 2 vectơ làn đường $\vec{l}_1$ và $\vec{r}_1$, sau đó tính góc nghiên của 2 vectơ này α_l và α_r theo công thức (3).

Bước 3: Chia vùng không gian còn lại của ảnh thành 4 phần (dựa trên độ dài của ảnh bản đồ làn đường) bằng cách sử dụng 3 đường quét ngang. Tiếp theo, thực hiện các bước xử lý sau:

Bước 3.1: Xác định vị trí của điểm $C_l^{'}$ dựa trên hướng của vectơ làn đường $\overrightarrow{l_2^{'}}$ là một sự mở rộng (nối dài) của vectơ làn đường $\overrightarrow{l_1}$ (Hình 2).

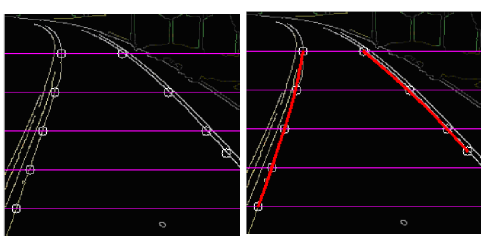
Bước 3.2: Xác định vị trí của điểm $C_r^{'}$ dựa trên hướng của vectơ làn đường $\overrightarrow{r_2^{'}}$ là một sự mở rộng (nối dài) của vectơ làn đường $\overrightarrow{r_1}$ (Hình 2).

Bước 3.3: Xác định vị trí trung điểm C_M của $C_l^{'}$ và $C_r^{'}$.

Bước 3.4: Quét từ điểm C_M sang trái để tìm điểm điều khiển C_l nằm trên làn đường bên trái. Nếu không tìm được C_l do làn đường bên trái không liên tục hoặc ảnh bị nhiễu ta sẽ sử dụng $C_l^{'}$ thay cho C_l.

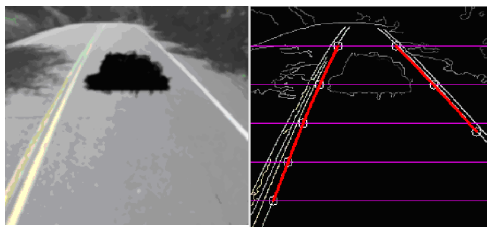
Bước 3.5: Quét từ điểm C_M sang phải để tìm điểm điều khiển C_r nằm trên làn đường bên phải. Nếu không tìm được C_r do làn đường bên phải không liên tục hoặc ảnh bị nhiễu ta sẽ sử dụng $C_r^{'}$ thay cho C_r.

Do chiều dài của các làn đường nhận từ ảnh chụp được từ camera thường là không bằng nhau, vì vậy chúng tôi chọn làn đường có chiều dài lớn hơn để phân chia phần không gian còn lại của ảnh thành 4 phần và tiến hành áp dụng giải thuật trên để tìm kiếm các điểm điều khiển. Quá trình này bảo đảm rằng ta có thể phát hiện được gần như đầy đủ làn đường (Hình 3). Ở đây chúng tôi chọn làn đường có độ dài dài hơn (làn đường bên trái) vì vậy điểm điều khiển đầu tiên không xuất hiện trên đường quét ngang ở làn đường bên phải.



Hình 3: Phát hiện làn đường sử dụng khái niệm vectơ-làn đường

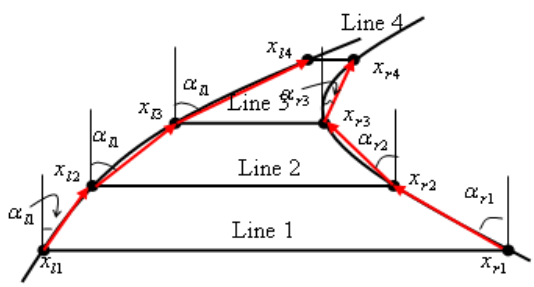
Trong một số trường hợp, khi chúng ta sử dụng đường quét từ vị trí trung điểm của đường quét sang trái hoặc sang phải để tìm các điểm điều khiển nằm trên các làn đường thì có thể xuất hiện nhiễu trên ảnh nằm ở vùng không gian giữa 2 làn đường (Hình 4 – trái). Để khắc phục điều này chúng tôi sử dụng một giá trị góc đặt trước DELTA_ANGLE. DELTA_ANGLE là giá trị khác biệt của giá trị 2 góc nghiêng liên tiếp trên làn đường trái hoặc phải. Nếu DELTA_ANGLE nhỏ hơn một giá trị ngưỡng định trước, ta tiếp tục di chuyển sang trái hoặc sang phải để tìm điểm điều khiển. Trong thử nghiệm của mình, chúng tôi thiết lập giá trị ngưỡng cho DELTA_ANGLE = 15^0. Một ví dụ minh họa được thể hiện như ở Hình 4.



Hình 4: Kết quả phát hiện làn đường với ảnh bị nhiễu giữa 2 làn đường

3 THUẬT TOÁN ƯỚC LƯỢNG ĐỘ CONG CỦA LÀN ĐƯỜNG

Sau khi phát hiện được giới hạn của các làn đường, việc tiếp theo chúng tôi sẽ rút trích thông tin để ước lượng độ cong hay hướng của các làn đường. Chúng tôi đề xuất thuật toán sau để ước lượng độ cong của các làn đường trái và phải một cách riêng biệt (Hình 5).



Hình 5: Mô hình ước lượng độ cong làn đường

Giải thuật ước lượng độ cong của các làn đường

Bước 1: Tính 3 góc ước lượng độ cong cho mỗi làn đường sử dụng 4 đường quét từ cuối ảnh như minh họa ở Hình 5 và công thức (4):

$$\begin{cases} \alpha_{li} = arctg\left(\dfrac{x_{l(i+1)} - x_{li}}{y_{l(i+1)} - y_{li}}\right) \\ \alpha_{ri} = arctg\left(\dfrac{x_{r(i+1)} - x_{ri}}{y_{r(i+1)} - y_{ri}}\right) \end{cases} \quad (i = 1..3) \qquad (4)$$

Bước 2: Tính giá trị khác biệt giữa các góc:

Bước 2.1: Đối với làn đường bên trái:

$$\Delta\alpha_{l1} = \alpha_{l2} - \alpha_{l1}; \quad \Delta\alpha_{l2} = \alpha_{l3} - \alpha_{l2} \quad (5)$$

Bước 2.2: Đối với làn đường bên phải:

$$\Delta\alpha_{r1} = \alpha_{r2} - \alpha_{r1}; \quad \Delta\alpha_{r2} = \alpha_{r3} - \alpha_{r2} \quad (6)$$

Bước 3: Tính giá trị trung bình (mean) của các góc:

Bước 3.1: Đối với làn đường bên trái:

$$\alpha_l = a_l\Delta\alpha_{l2} + b_l\Delta\alpha_{l1} + c_l\alpha_{l1} \quad (7)$$

Bước 3.2: Đối với làn đường bên phải:

$$\alpha_l = a_l\Delta\alpha_{l2} + b_l\Delta\alpha_{l1} + c_l\alpha_{l1} \quad (8)$$

Bước 3.3: Chọn dấu cho các hệ số a_l, b_l, c_l dựa vào các công thức (9a), (9b), (9c):

$$\begin{cases} a_l \geq 0 \; if \; \Delta\alpha_{l2} \geq 0 \\ a_l < 0 \; if \; \Delta\alpha_{l2} < 0 \end{cases} \quad (9a)$$

$$\begin{cases} b_l \geq 0 \; if \; \Delta\alpha_{l1} \geq 0 \\ b_l < 0 \; if \; \Delta\alpha_{l1} < 0 \end{cases} \quad (9b)$$

$$\begin{cases} c_l \geq 0 \; if \; \alpha_{l1} \geq 0 \\ c_l < 0 \; if \; \alpha_{l1} < 0 \end{cases} \quad (9c)$$

Bước 3.4: Chọn dấu cho các hệ số a_r, b_r, c_r dựa vào các công thức (10a), (10b) và (10c):

$$\begin{cases} a_r \geq 0 \; if \; \Delta\alpha_{r2} \geq 0 \\ a_r < 0 \; if \; \Delta\alpha_{r2} < 0 \end{cases} \quad (10a)$$

$$\begin{cases} b_r \geq 0 \; if \; \Delta\alpha_{r1} \geq 0 \\ b_r < 0 \; if \; \Delta\alpha_{r1} < 0 \end{cases} \quad (10b)$$

$$\begin{cases} c_r \geq 0 \; if \; \alpha_{r1} \geq 0 \\ c_r < 0 \; if \; \alpha_{r1} < 0 \end{cases} \quad (10c)$$

Giá trị độ lớn của a_l, b_l, c_l và a_r, b_r, c_r được thiết lập và lựa chọn dựa vào thực nghiệm. Trong thử nghiệm của chúng tôi thì $|a_l| = |a_r| = 3, |b_l| = |b_r| = 2, |c_l| = |c_r| = 1$.

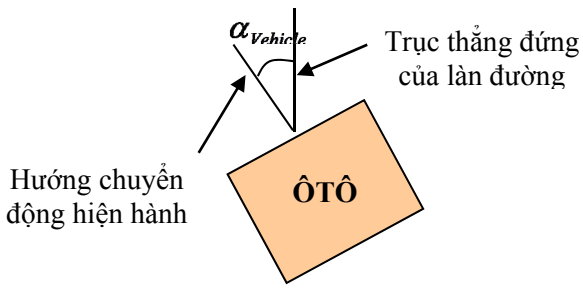
4 TÍNH TOÁN GÓC LÁI VÀ LẦN VẾT LÀN ĐƯỜNG CHO ÔTÔ TỰ HÀNH

Để lần vết làn đường và điều khiển góc lái cho ôtô tự hành chuyển động dọc theo làn đường quy định thì góc lái của xe cần phải được xác định. Góc lái của xe có thể được tính toán dựa trên các điều kiện:

– Độ cong của làn đường trái α_l và làn đường phải α_r

– Vị trí hiện hành của ôtô trên làn đường, ký hiệu là $p_{Vehicle}$. Ở đây $p_{Vehicle}$ phải thỏa mãn điều kiện $0 \leq p_{Vehicle} \leq W$ với W là độ rộng của làn đường. Trong thực nghiệm của chúng tôi $W = 256$ là độ rộng của ảnh nhận được từ camera.

– Hướng di chuyển hiện hành của ôtô, ký hiệu là $\alpha_{Vehicle}$. Ở đây $\alpha_{Vehicle}$ là góc tạo bởi thân của ôtô và trục thẳng đứng như minh họa ở Hình 6.



Hình 6: Minh họa vị trí ôtô trong làn đường

Dựa trên các tham số đã định nghĩa trên, chúng tôi tiến hành xác định giá trị của góc lái cần thiết để điều khiển ôtô tự hành có thể chuyển động trong làn đường quy định.

Giải thuật tính toán góc lái của ôtô

Bước 1: Tính toán 2 hệ số k_l và k_r tương ứng lần lượt cho các làn đường trái và phải sử dụng các công thức (11) sau:

$$k_l = 2\frac{(W - p_{Vehicle})}{W}; \; k_r = 2(1 - k_l) \quad (11)$$

Bước 2: Tính toán giá trị đầu ra của góc lái như sau:

Nếu α_l và α_r trái dấu nhau:

$$\alpha_{Out} = \frac{(k_l\alpha_l + k_r\alpha_r)}{(k_l + k_r)} - \alpha_{Vehicle} \quad (12)$$

Ngược lại, nếu α_l và α_r cùng dấu nhau:

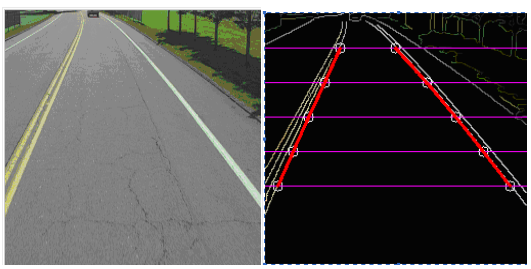
$$\alpha_{Out} = \begin{cases} \frac{(k_l\alpha_l - k_r\alpha_r)}{(k_l + k_r)} - \alpha_{Vehicle} & if \; |\alpha_l \geq \alpha_r| \\ \frac{(k_r\alpha_r - k_l\alpha_l)}{(k_l + k_r)} - \alpha_{Vehicle} & if \; |\alpha_l < \alpha_r| \end{cases} \quad (13)$$

5 MỘT SỐ KẾT QUẢ THỰC NGHIỆM

5.1 Kết quả thực nghiệm minh họa giải thuật phát hiện làn đường

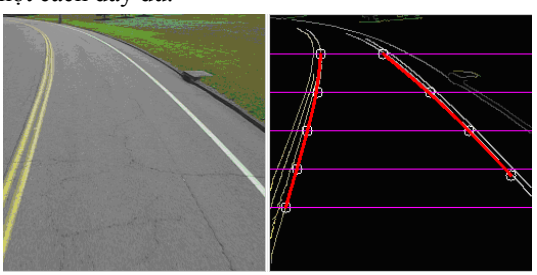
Giải thuật phát hiện làn đường được kiểm tra trên tập ảnh sưu tập từ camera trong các điều kiện khác nhau về thời gian và vị trí chụp ảnh. Hình 7 trình bày một số kết quả kiểm tra giải thuật được đề xuất trong một số trường hợp tiêu biểu sử dụng khái niệm vectơ-làn đường.

Trường hợp 1: Làn đường có dạng gần như đường thẳng, không có nhiễu và độ dài của 2 làn đường trái và phải là xấp xỉ nhau. Trong trường hợp này, giải thuật của chúng tôi phát hiện làn đường rất tốt và xây dựng lại đầy đủ làn đường bằng cách sử dụng hàm nội suy NUBS bậc 3 như minh họa sau:

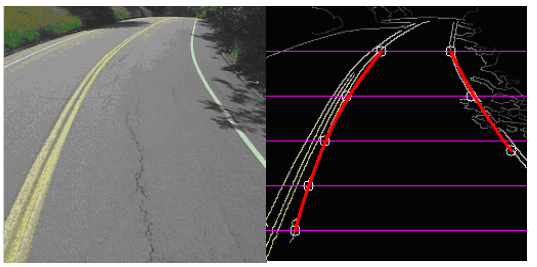


Trường hợp 2: Làn đường có dạng cong lệch sang trái hoặc sang phải và ảnh bị nhiễu bởi bóng cây trên đường. Trong trường hợp này, hàm nội suy NBUS bậc 2 hoặc bậc 3 sẽ được sử dụng để xây dựng lại làn đường tùy thuộc vào số lượng điểm điều khiển được phát hiện nhờ kỹ thuật vectơ làn đường.

Trường hợp 2.1: Làn đường có dạng cong lệch sang trái một ít và ảnh thu được không có nhiễu. Trong trường hợp này, số điểm điều khiển được phát hiện trên 2 làn đường là không bằng nhau. Tuy nhiên, làn đường vẫn được xây dựng lại một cách đầy đủ.

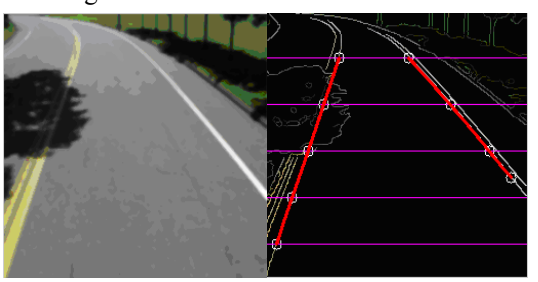


Trường hợp 2.2: Làn đường có dạng cong lệch sang nhiều sang phải và ảnh thu được có nhiễu bóng cây trên làn đường phải. Trong trường hợp này, sử dụng vectơ-làn đường giúp ta vượt qua vấn đề khó khăn về nhiễu. Làn đường được phát hiện và xây dựng lại đầy đủ.

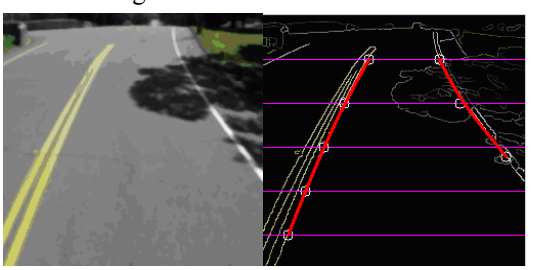


Trường hợp 3: Làn đường có dạng cong lệch sang trái hoặc sang phải và ảnh bị nhiễu khá nhiều bởi bóng cây hay trụ đèn trên đường. Trong các trường hợp này, nếu không tìm được các điểm điều khiển chính xác, chúng tôi sử dụng vectơ làn đường để tính toán vị trí điểm điều khiển và dùng hàm nội suy NBUS bậc 2 hoặc bậc 3 sẽ được sử dụng để xây dựng lại làn đường tùy thuộc vào số lượng điểm điều khiển được tính toán nhờ kỹ thuật vectơ làn đường.

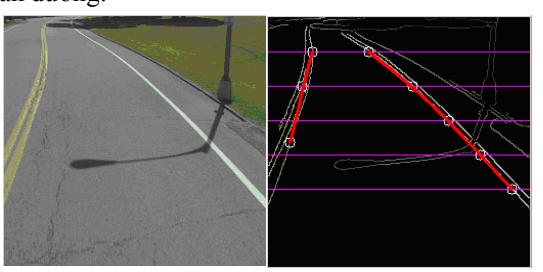
Trường hợp 3.1: Làn đường có dạng cong lệch sang trái và nhiễu do bóng cây xuất hiện nhiều trên làn đường:



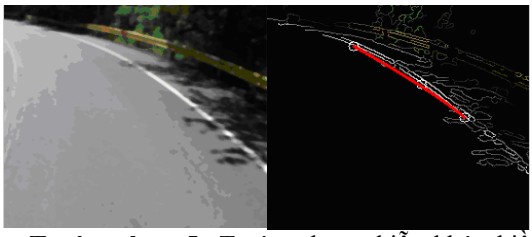
Trường hợp 3.2: Làn đường có dạng cong lệch sang phải và nhiễu do bóng cây xuất hiện khá dày trên làn đường:



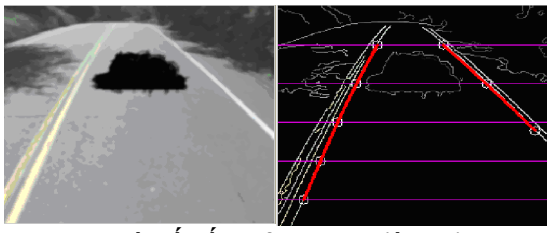
Trường hợp 3.3: Làn đường có dạng cong lệch sang trái và nhiễu do bóng cột đèn hiện nhiều trên làn đường:



Trường hợp 4: Làn đường có dạng cong lệch nhiều sang một phía và làn đường còn lại bị mất (ảnh do camera chụp không nhìn thấy được làn đường còn lại). Ngoài ra, làn đường còn bị nhiễu rất dày nhưng giải thuật của chúng tôi vẫn phát hiện và xây dựng lại thành công dạng làn đường này.



Trường hợp 5: Trường hợp nhiễu khá nhiều xuất hiện trong vùng giới hạn của 2 làn đường giải thuật của chúng tôi định giá trị ngưỡng của DELTA_ANGLE kết hợp với vectơ làn đường cho phép xây dựng được các điểm điều khiển cho phép phát hiện và xây dựng lại các làn đường một cách chính xác.

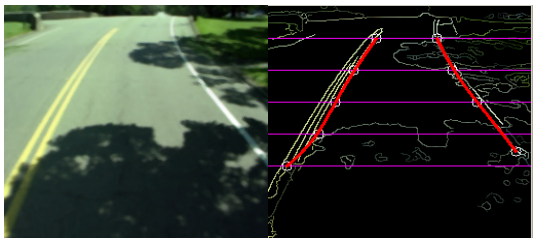


Hình 7: Một số kết quả thực nghiệm minh họa giải thuật phát hiện làn đường

Trong trường hợp ảnh bị nhiễu quá nặng và nhiễu che khuất hầu hết vết sơn đánh dấu làn đường thì giải thuật phát hiện làn đường của chúng tôi bị thất bại không xây dựng lại được làn đường (Hình 8).

Trường hợp 1: Làn đường có dạng cong lệch nhiều sang phải và vết sơn đánh dấu làn đường bị nhiễu bóng cây che phủ gần như toàn bộ. Trong trường hợp này, giải thuật của chúng tôi không xây dựng được vectơ làn đường cho làn đường bên phải và không phát hiện được làn đường bên phải này.

Trường hợp 2: Làn đường có dạng cong lệch nhiều sang phải và nhiễu do bóng cây che phủ phần lớn lòng đường và một phần làn đường. Trong trường hợp này, đôi khi giải thuật của chúng tôi phát hiện và xây dựng được làn đường nhưng không chính xác hoàn toàn. Trong đa phần các trường hợp có dạng như thế này giải thuật của chúng tôi cũng bị thất bại.



Hình 8: Các trường hợp giải thuật phát hiện làn đường bị thất bại

Chúng tôi sử dụng 2 tập dữ liệu ảnh từ cơ sở dữ liệu ảnh [10] để chạy kiểm tra giải thuật được đề xuất. Tập dữ liệu 1 chứa 69 ảnh và tập dữ liệu 2 chứa 99 ảnh chụp các dạng làn đường khác nhau ở các thời điểm khác nhau. Đối với tập ảnh đầu tiên, giải thuật của chúng tôi phát hiện làn đường thất bại trong 5 trường hợp và với tập ảnh thứ 2 là 7 trường hợp. Thời gian thực hiện giải thuật là hơn 13 frames ảnh/giây. Nói tóm lại, phương pháp của chúng tôi phát hiện một làn đường thất bại nếu không thể xác định được 2 điểm điều khiển đầu tiên để xây dựng vectơ làn đường cho làn đường này.

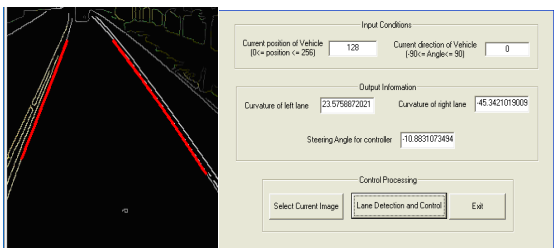
5.2 Kết quả thực nghiệm ước lượng độ cong làn đường và xác định góc lái

Trong phần này, chúng tôi trình bày một số kết quả mô phỏng minh họa cho giải thuật ước lượng độ cong của làn đường tính giá trị đầu ra của góc điều khiển lái cho ôtô tự hành.

Trường hợp 1: $P_{Vehicle} = \frac{W}{2}$ và $\alpha_{Vehicle} = 0^0$

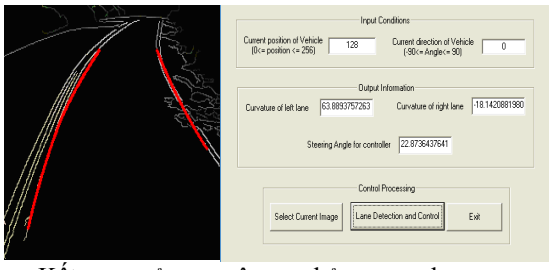
Trong trường hợp này, vị trí của ôtô là chính giữa hai làn đường, và góc lái hiện hành của ôtô là 0. Bốn dạng khác nhau của trường hợp này được tiến hành thử nghiệm và mô phỏng minh họa như ở các hình tiếp theo.

Trường hợp 1.1: Các làn đường có dạng gần như đường thẳng và không bị nhiễu



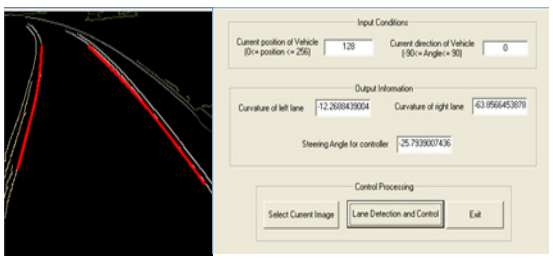
Kết quả mô phỏng như sau: $\alpha_l = 23.576^0$; $\alpha_r = -45.342^0 \Rightarrow \alpha_{Out} = -10.883^0$. Điều này cho biết từ vị trí này, ôtô sẽ quẹo một ít sang trái với góc lái là 10.883^0.

Trường hợp 1.2: Làn đường có dạng cong sang phải và làn đường bị nhiễu một ít.



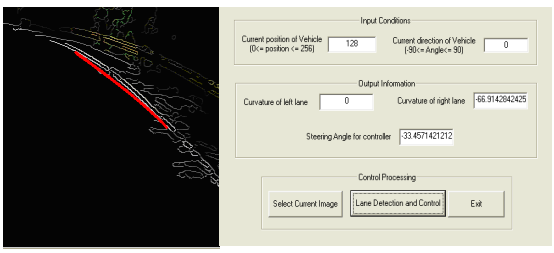
Kết quả mô phỏng như sau: $\alpha_l = 63.889^0$; $\alpha_r = -18.142^0 \Rightarrow \alpha_{Out} = 22.874^0$. Điều này cho biết từ vị trí này, ôtô sẽ quẹo sang phải với một góc lái trung bình và có giá trị là 22.874^0

Trường hợp 1.3: Làn đường có dạng cong sang trái và không bị nhiễu.



Kết quả mô phỏng như sau: $\alpha_l = -12.269^0$; $\alpha_r = -63.857^0 \Rightarrow \alpha_{Out} = -25.794^0$. Điều này cho biết từ vị trí này, ôtô sẽ quẹo sang trái với một góc lái trung bình có giá trị là 25.749^0

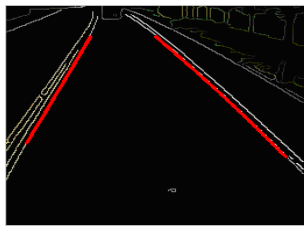
Trường hợp 1.4: Làn đường bên phải có dạng cong rất nhiều sang trái, làn đường bên phải và làn đường bên trái bị mất.



Kết quả mô phỏng như sau: $\alpha_l = 0^0$; $\alpha_r = -66.914^0 \Rightarrow \alpha_{Out} = -33.457^0$. Điều này cho biết từ vị trí này, ôtô sẽ quẹo sang trái với một góc lái tương đối lớn có giá trị 33.457^0

Trường hợp 2: $0 \le P_{Vehicle} \le W$ và $\alpha_{Vehicle} = 0^0$

Với trường hợp này, ôtô đang ở vị trí bất kỳ giữa hai làn đường, và góc lái hiện hành của ôtô là 0. Một số dạng của trường hợp này được mô phỏng với dạng làn đường như sau:



Trường hợp 2.1: $P_{Vehicle} = \frac{1}{4}W$

$\Rightarrow \alpha_{Out} = 6.346^0$

Kết quả mô phỏng cho thấy rằng ôtô cần thay đổi chuyển động một ít sang phải từ vị trí hiện tại. Kết quả này cũng rất phù hợp vì vị trí hiện hành của ôtô là ở gần làn đường giới hạn trái và làn đường chỉ cong sang trái một ít.

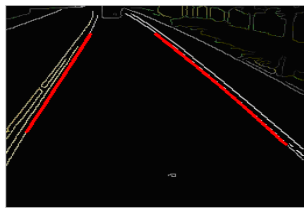
Trường hợp 2.2: $\frac{3}{4}W < P_{Vehicle} < W$

$\Rightarrow \alpha_{Out} = -30.266^0$

Kết quả mô phỏng cho thấy rằng ôtô cần thay đổi chuyển động sang trái với một góc lái lớn từ vị trí hiện tại. Kết quả này cũng rất phù hợp vì vị trí hiện hành của ôtô là ở gần làn đường giới hạn phải và làn đường lại cong sang trái một ít.

Trường hợp 3: $0 \le P_{Vehicle} \le W$ và $\alpha_{Vehicle} \ne 0^0$

Ở trường hợp này, ôtô được xét đang ở vị trí bất kỳ giữa hai làn đường, và góc lái hiện hành của ôtô là khác 0. Một số dạng của trường hợp này được mô phỏng với dạng làn đường sau:



Trường hợp 3.1: $P_{Vehicle} = \frac{1}{2}W, \alpha_{Vehicle} = 10^0$

$\Rightarrow \alpha_{Out} = -20.883^0$

Kết quả mô phỏng cho thấy rằng ôtô cần thay đổi chuyển động sang trái từ vị trí hiện tại với góc lái tương đối có độ lớn là 20.883^0.

Trường hợp 3.2: $P_{Vehicle} = \frac{W}{4}, \alpha_{Vehicle} = -30.5^0$

$\Rightarrow \alpha_{Out} = -6.154^0$

Kết quả mô phỏng cho thấy rằng ôtô cần thay đổi chuyển động sang trái từ vị trí hiện tại với góc lái nhỏ có độ lớn là 6.154^0.

6 KẾT LUẬN VÀ ĐỀ XUẤT

Bài báo đã đề xuất một giải thuật xử lý ảnh mới và đơn giản cho bài toán phát hiện, xây dựng làn đường và điều khiển góc lái cho ôtô tự hành có thể chuyển động trong phạm vi làn đường cho phép. Tốc độ xử lý của hệ thống là hơn 13 frames ảnh/giây. Mặc dù không có điều kiện thực hiện các hệ thống ôtô tự hành kết hợp sensor hoặc camera với bộ điều khiển fuzzy hay neural nhưng có thể khẳng định chắc chắn rằng hệ thống được đề xuất là đơn giản hơn và cho đáp ứng nhanh hơn các hệ thống kết hợp bộ điều khiển fuzzy hay neural. Lý do là vì thông tin nhận được qua các kỹ thuật xử lý ảnh đơn giản sẽ được hệ thống tính ra góc lái tức thời không cần thời gian chờ đợi đáp ứng của bộ điều khiển mà thông thường các bộ điều khiển đều cần có một thời gian trễ nhất định để phản hồi lại thông tin cho hệ thống (tối thiểu cũng xấp xỉ 0.1s). Các kết quả thử nghiệm chứng minh rằng phương pháp đề nghị là phù hợp cho việc điều khiển ôtô chuyển động chính xác trong giới hạn làn đường cho phép mà không cần sự can thiệp của con người. Một số điểm mạnh của hệ có thể được tóm tắt như sau:

– Phương pháp phát hiện làn đường có thể được áp dụng cho việc nhận dạng và xây dựng nhiều dạng làn đường khác nhau và có độ cong lớn dựa trên khái niệm vectơ-làn đường.

– Làn đường trái và phải được phát hiện và xây dựng tách biệt không giống như một số nghiên cứu đã có dựa trên giả thuyết 2 làn đường là đồng nhất.

– Kỹ thuật sử dụng vectơ-làn đường cho phép vượt qua các khó khăn về nhiễu trên ảnh để phát hiện đầy đủ các làn đường không liên tục, hoặc bị nhiễu do các đối tượng khác tạo ra.

– Ước lượng độ cong của các làn đường để cho phép điều khiển góc lái của ôtô chuyển động theo đúng giới hạn làn đường cho phép.

Trong thời gian tới, nếu điều kiện cho phép chúng tôi sẽ xây dựng mô hình ôtô tự hành và kiểm nghiệm giải thuật được đề nghị trong môi trường thực cũng như xây dựng giải thuật cho phép phát hiện và tránh vật cản trên đường di chuyển của robot hay ôtô.

TÀI LIỆU THAM KHẢO

1. M.Asif, M.R.Arshad, and P.A.Wilson, 2007. "AGV Guidance System: An Application of Simple Active Contour for Visual Tracking", Proceeding world academy of science engineering and technology, Vol. 6, No. 2, pp. 664 –667.
2. Yue Wang, Eam Khwang Teoh, Dinggang Shen, 2004. "Lane detection and tracking using Bsnake", Journal of Image and Vision Computing, Vol. 22, No. 1, pp. 269–280.
3. Jiang Ruyi, Klette Reinhard, Vaudrey Tobi, Wang Shigang, 2011. "Lane detection and tracking using a new lane model and distance transform", Journal of Machine Vision and Application, Vol. 22, No. 4, pp. 721–737.
4. Mohamed Aly, 2008. "Real time Detection of Lane Markers in Urban Streets", International Conference on Intelligent Vehicles Symposium, pp.7-12.
5. Hodge, N. E., Shi, L. Z., Trabia, M. B., 2004. "A distributed Fuzzy Logic Controller for an Autonomous Vehicle. Journal of Robotic and Systems, Vol. 21, No.10, 499-516.
6. Castro, A. P. A., da Silva, J. D. S., Simoni, P. O., 2001. "Image based Autonomous Navigation with Fuzzy Logic Control". Proceeding on IEEE International conference on Neural Network, pp.2200-2205.
7. Ryoo, Y. J., Lim, Y. C., 1999. "Neuro-Fuzzy Control system for Vision-Based Autonomous Vehicle", Proceeding on IEEE International conference on Fuzzy system, pp.1643-1648.
8. Lu, H. E., Wang, P. S. P., 1986. "A Comment on "a Fast Parallel Algorithm for Thinning Digital Patterns", Journal o Communication of the ACM, Vol. 29, No. 6, pp. 239-242.
9. Chihab, N., Zergainoh, A., Astruc, J.-P., 2003. "Generalized non-uniform B-spline functions for discrete signal interpolation", Proceeding on IEEE International symposium on Signal processing and its applications, pp.129-132.
10. http://vasc.ri.cmu.edu//idb/html/road/may30_90/index.html.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

THEO DÕI HIỆN TRẠNG TRÀ LÚA PHỤC VỤ CẢNH BÁO DỊCH HẠI LÚA TRÊN CƠ SỞ SỬ DỤNG CÔNG NGHỆ VIỄN THÁM VÀ HỆ THỐNG THÔNG TIN ĐỊA LÝ GIS

Trần Thị Hiền[1], Võ Quang Minh[2], Huỳnh Thị Thu Hương[2], Trần Thanh Dân[2], Hồ Văn Chiến[3], Nguyễn Hữu An[4] và Nguyễn Phước Thành[4]

[1] *Sở Tài nguyên và Môi trường tỉnh Bến Tre*
[2] *Bộ môn Tài nguyên Đất đai, Khoa Môi trường & Tài nguyên Thiên nhiên, Trường Đại học Cần Thơ*
[3] *Trung Tâm Bảo vệ thực vật Phía Nam*
[4] *Chi cục Bảo vệ thực vật An Giang*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Rice crop monitoring for early warning pest occurence using remote sensing and geographic information systems

Từ khóa:
Trà lúa, dịch hại, mùa vụ, viễn thám, MODIS, NDVI

Keywords:
Cropping stages, season, pest, remote sensing, MODIS, NDVI

ABSTRACT

MODIS (MODerate Resolution Spectroradiometer) Remote Sensing Technology with varying multispectral spatial and temporal resolution and large coverage, which suitable for monitoring the changes of Earth's surface in large scale. Some kind of data from satelite images provides information to monitor vegetation cover such as MOD09Q1 image, with spatial resolution of 250x250m, and at 8-day interval, suitable for the study of crops and rice cropping calendar. Based on the images properties (Normalized Difference Vegetatione Index-NDVI), the relationship with the change of rice crop status in space and time to help identify the rice cropping calendar and cropping season (case study in An Giang province 2012-2013). Besides, the the proediction of pest occurences could be generated based on the correlation between the rice growing stages and pests occurence. Results show that can use satellite images of MODIS-MOD09Q1 to generate the rice cropping stages and pest occurrens status. It has a close correlation between vegetation indices with different rice growing stages at high accuracy (R^2 = 0.83).

TÓM TẮT

Công nghệ ảnh viễn thám MODIS (Moderate Resolution Spectroradiometer) đa phổ với nhiều mức độ phân giải không gian và thời gian khác nhau và có độ phủ lớn thích hợp cho việc theo dõi biến động thông tin trên bề mặt trái đất ở phạm vi lớn. Trong đó có một số loại ảnh chuyên cung cấp thông tin để theo dõi độ phủ thực vật như ảnh MOD09Q1 độ phân giải không gian 250x250m, chu kỳ 8 ngày lập, thích hợp cho nghiên cứu các loại thực vật canh tác theo mùa vụ đặc biệt là cây lúa –loại cây trồng chủ yếu. Dựa vào đặc điểm ảnh (giá trị chỉ số khác biệt thực vật NDVI) có mối quan hệ với sự thay đổi của hiện trạng sinh trưởng của cây lúa theo không gian và thời gian giúp xác định thời gian xuống giống, biến động không gian hiện trạng trà lúa và cơ cấu mùa vụ các vùng trồng lúa (trường hợp nghiên cứu ở tỉnh An Giang năm 2012-2013). Bên cạnh đó, hệ thống bản đồ cảnh báo dịch hại được xây dựng dựa trên mối tương quan giữa hiện trạng trà lúa và các loại dịch hại, côn trùng tấn công. Kết quả cho thấy có thể sử dụng ảnh vệ tinh độ phân giải thấp đa thời gian MODIS-MOD09Q1 để xây dựng bản đồ hiện trạng trà lúa. Nó có mối liên hệ chặt chẽ giữa chỉ số khác biệt thực vật với các giai đoạn tăng trưởng của cây. Đồng thời qua kết quả kiểm tra, đối chiếu cho thấy kết quả giải đoán có độ chính xác cao với độ tin cậy R^2=0,83. Do đó, có thể ứng dụng ảnh MODIS độ phân giải thấp này để xác định hiện trạng trà lúa dựa trên mối tương quan giữa chỉ số NDVI và các giai đoạn phát triển của cây lúa.

1 GIỚI THIỆU

Hiện nay, cơ cấu mùa vụ lúa ở đồng bằng sông Cửu Long (ĐBSCL) rất phức tạp khiến cho công tác điều tra, thống kê, theo dõi thời vụ xuống giống cũng như lập bản đồ hiện trạng vùng trồng lúa trở nên khó khăn. Công tác theo dõi thời vụ xuống giống chủ yếu dựa vào điều tra, các báo cáo định kỳ. Điều này không đảm bảo độ chính xác, đòi hỏi nhiều thời gian, kinh phí và thường không đáp ứng kịp nhu cầu ra quyết định, hoạch định chính sách của địa phương. Vì thế đòi hỏi người nông dân cần có kiến thức tổng quát về qui trình thâm canh tổng hợp, hiểu rõ về quy luật phát sinh phát triển của các loài sâu bệnh (Phạm Văn Dư và Lê Cầm Loan, 2007). Trong khi đó dịch hại trên lúa có liên quan đến thời vụ xuống giống, loại giống, giai đoạn sinh trưởng của cây lúa, đặc biệt ở giai đoạn trước 20 ngày tuổi. Do đó, việc theo dõi được diện tích gieo trồng của các giống lúa và trà lúa ở từng thời điểm của từng địa phương sẽ giúp cho các nhà quản lý nông nghiệp đặc biệt ngành Bảo vệ thực vật dễ dàng quản lý chỉ đạo sản xuất và dự báo sớm tình hình dịch hại trên lúa ở từng cấp quản lý, cũng như theo dõi tiến trình áp dụng các biện pháp quản lý, bảo vệ cây trồng ở từng địa phương.

Ngày nay việc sử dụng kỹ thuật ảnh viễn thám, kết hợp với kỹ thuật thông tin địa lý (GIS) phương pháp thống kê và nội suy không gian đã được ứng dụng trong nghiên cứu phân bố không gian các đặc tính tự nhiên ở nhiều nước trên thế giới (Burrough, P. A, 1986, Aronoff S. 1989). MODIS (Moderate Resolution Imaging Spectroradiometer) là bộ cảm đặt trên vệ tinh TERRA và vệ tinh AQUA. Đây là hai vệ tinh nghiên cứu môi trường của NASA (Hoa Kỳ), cung cấp các loại ảnh với mục đích quan trắc, theo dõi các thông tin về mặt đất, đại dương và khí quyển trên phạm vi toàn cầu. Trong phạm vi nghiên cứu đã sử dụng ảnh MOD13Q1 được thu thập miễn phí từ vệ tinh quan trắc trái đất Terra của NASA tại địa chỉ https://wist.echo.nasa.gov. Đây là một sản phẩm điển hình của dòng MODIS Level 3 được sử dụng chủ yếu để tính toán chỉ số NDVI trong dự báo khô hạn, hiện trạng cơ cấu mùa vụ,… Kết quả có thể giải đoán, theo dõi xác định được thời điểm và tiến độ xuống giống ở các địa phương sẽ đánh giá được hiện trạng canh tác lúa và dự đoán những vùng có nguy cơ dịch hại làm cơ sở đề xuất các giải pháp phù hợp với từng trà lúa ở từng vùng khác nhau.

Nghiên cứu chủ yếu sử dụng ảnh viễn thám MODIS kết hợp kỹ thuật GIS để theo dõi tiến độ xuống lúa và hiện trạng trà lúa. Kết quả được so sánh kiểm chứng với các dữ liệu điều tra hiện trạng và tình hình dịch hại lúa trên toàn vùng của Trung Tâm Bảo vệ thực vật Phía Nam.

2 KHU VỰC NGHIÊN CỨU

ĐBSCL nằm ở vùng cực Nam của nước Việt Nam, nằm trong vùng từ $8^0$30'-11^0 vĩ độ Bắc và từ $104^0$30'-107^0 kinh độ Đông, gồm 13 tỉnh: Long An, Tiền Giang, Bến Tre, Đồng Tháp, Vĩnh Long, An Giang, Cần Thơ, Hậu Giang, Sóc Trăng, Cà Mau, Bạc Liêu, Kiên Giang, Trà Vinh (Lê Sâm, 1996).

Đặc điểm ảnh

Ảnh được thu thập từ cơ quan NASA (Mỹ) có ký hiệu MOD13Q1 với có độ phân giải không gian 250 m, độ phân giải thời gian với 16 ngày lập, với tổng số pixel là 4800 x 4800, hệ tọa độ kinh độ, vĩ độ (lat/long). Trong phạm vi đề tài ảnh NDVI được sử dụng để phục vụ cho quá trình giải đoán. Ảnh NDVI là ảnh được tính toán được dựa vào hai dãy phổ đỏ (Red) và cận hồng ngoại (NIR) (Bảng 1).

Bảng 1: Các kênh phổ của đầu đo MODIS sử dụng trong tính toán chỉ số thực vật

Kênh MODIS	Bước sóng (*m*µ)	Độ rộng bước sóng (*m*µ)	Độ phân giải (m)
1	0,620-0,670	0,005	250
2	0,841-0,876	0,035	250

3 PHƯƠNG PHÁP

3.1 Tạo ảnh chỉ số thực vật

Chỉ số phổ thực vật được kết hợp từ các band phổ nhìn thấy, cận hồng ngoại, hồng ngoại và đỏ là các tham số trung gian mà từ đó có thể đánh giá được các đặc tính khác nhau của thực vật như: sinh khối, chỉ số diện tích lá, khả năng quang hợp theo mùa. Những đặc tính đó có liên quan và phụ thuộc rất nhiều vào dạng thực vật bao phủ và thời tiết, đặc tính sinh lý, sinh hoá và sâu bệnh.

Chỉ số khác biệt thực vật (NDVI) được trung bình hoá trong một chuỗi số liệu theo thời gian sẽ là công cụ cơ bản để giám sát sự thay đổi trạng thái lớp phủ thực vật, trên cơ sở đó biết được tác động của thời tiết, khí hậu đến sinh quyển. Chỉ số thực vật NDVI được tính theo công thức sau:

$$NDVI = \frac{(NIR - R)}{(NIR + R)}$$

3.2 Mối quan hệ giữa giá trị NDVI và giai đoạn phát triển của cây lúa

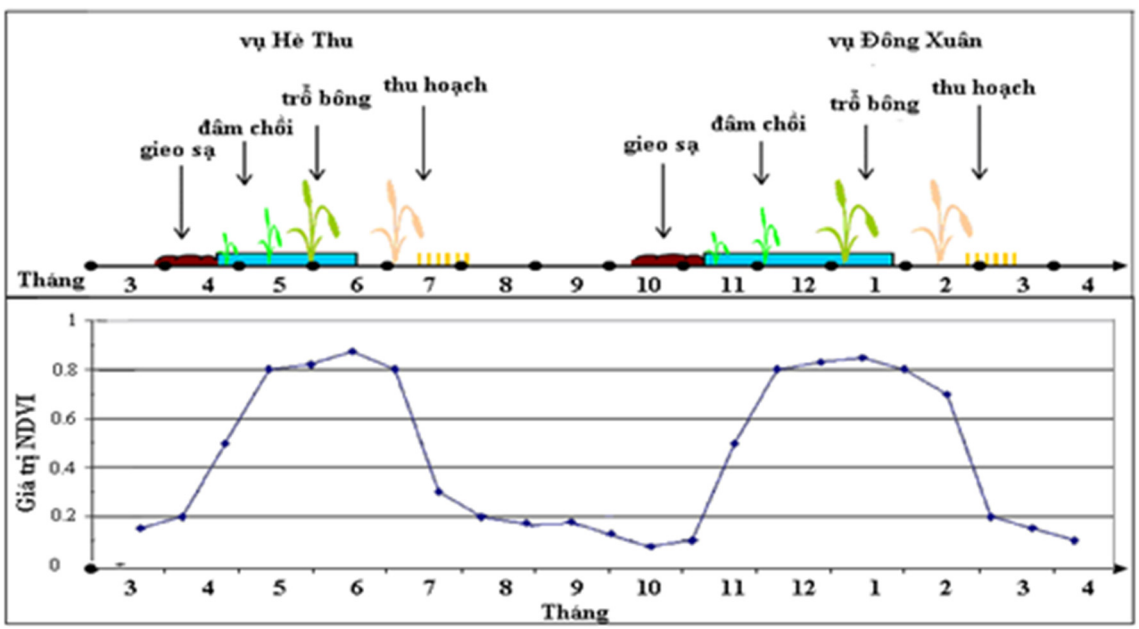
Mỗi loại cây trồng có khoảng giá trị NDVI dao động trong một khoảng giới hạn nhất định (do trên mỗi loại đất có đặc tính khác nhau, trên những vùng đất màu mỡ thì cây trồng phát triển tốt giái trị NDVI sẽ đạt cao và ngược lại), nhìn chung quy luật biến động của chúng giống nhau. Dựa trên kết quả nghiên cứu của B.R. Parida *et al* (2008) ở bang Gujarat của Ấn Độ đã đề xuất khoảng dao động NDVI cho một số đối tượng sau (Bảng 2).

Bảng 2: Tiêu chuẩn sử dụng để phân loại sử dụng đất (B.R. Parida *et al.*, 2008)

Khoảng biến động	Loại
NDVI > 0,74	Rừng (forest)
0,74 < NDVI > 0,46	Cây mùa vụ có tưới (Irrigated crops)
0,46 < NDVI > 0,20	Cây mùa vụ nước trời (Rainfed crops)
0,20 < NDVI > 0,15	Đất hoang (Fallow land)
0,15 < NDVI > 0,05	Đất trống (Bare soils)
0,05 < NDVI > 0,001	Đất làm muối (Salt pans)
NDVI < -0,001	Nước (Water)

3.3 Theo dõi tiến độ xuống giống và thành lập bản đồ phân bố thời vụ canh tác

Từ kết quả xử lý, kết hợp khảo sát thực tế, thống kê chỉ số NDVI theo thời gian của toàn vùng ĐBSCL, đối chiếu với hiện trạng sinh trưởng của cây lúa thực tế tìm ra mối quan hệ giữa chỉ số NDVI với hiện trạng các trà lúa. Trên cơ sở lý thuyết chỉ số NDVI của vùng lúa ở ĐBSCL có giá trị thấp vào đầu vụ, tăng dần và đạt cao nhất vào lúc cây lúa ở giai đoạn sau khi đẻ nhánh, giảm khi cây lúa bắt đầu chín và thấp nhất vào cuối vụ (Hình 1).



Hình 1: Sự biến đổi của chỉ số khác biệt thực vật ở các giai đoạn phát triển của cây lúa vụ Đông xuân - Hè Thu (Trần Thị Hiền và Võ Quang Minh, 2010)

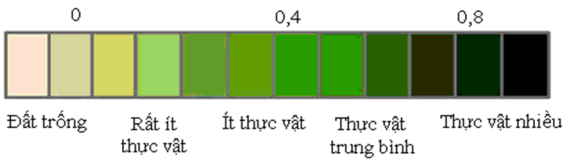
3.4 Dự đoán sự xuất hiện của dịch hại

Từ kết quả xác định hiện trạng các trà lúa ở từng thời điểm khảo sát và dựa vào khả năng xuất hiện và tấn công của các loại dịch hại tương ứng ở từng trà lúa khác nhau, bản đồ dự đoán sự xuất hiện của các loại dịch hại được thành lập, trên cơ sở sử dụng kỹ thuật GIS. Kết quả dự đoán cũng được so sánh với các số liệu báo cáo của cơ quan chuyên ngành.

4 KẾT QUẢ VÀ THẢO LUẬN

4.1 Giải đoán ảnh NDVI đa thời gian

Mỗi ảnh NDVI cho phép ta giám sát sự thay đổi trạng thái lớp phủ thực vật, thể hiện qua sự khác biệt của chỉ số NDVI ở một thời điểm nhất định, thể hiện mức độ đậm nhạt tương tứng với các giá trị NDVI như ở Hình 2

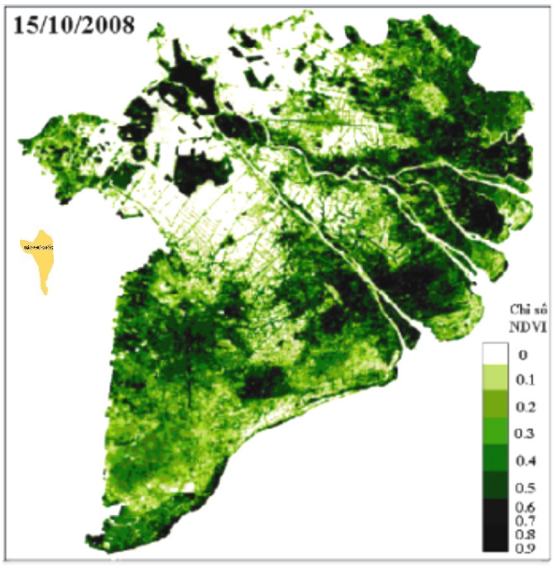


Hình 2: Mối quan hệ giữa chỉ số NDVI với sự hiện diện của thực vật

Qua phân tích ảnh NDVI năm 2008, một vài nơi sắc độ ảnh ổn định (sáng hoặc tối) quanh năm, còn lại đều biến động theo thời gian. Khảo sát 6 tháng đầu năm 2008 cho thấy những vùng ven biển và bán đảo Cà Mau có chỉ số NDVI rất thấp (sắc độ sáng) và không thay đổi rõ rệt trong năm. Ngược lại, vùng An Giang, Đồng Tháp, Cần Thơ, Hậu Giang có giá trị NDVI thay đổi, cao vào tháng 1, 2 giảm dần vào tháng 3, 4 và tăng trở lại vào tháng 5, 6. Trong 6 tháng cuối năm 2008 một phần bán đảo Cà Mau và ven biển, giá trị NDVI sắc độ ảnh có xu hướng đậm dần theo thời gian (NDVI giảm dần). Ngược lại, ở các tỉnh An Giang, Đồng Tháp và một phần tỉnh Kiên Giang thì giá trị NDVI lại giảm vào các tháng 9, 10 và 11. Kết quả phân tích các ảnh năm 2009 cũng cho thấy có sự biến động tương tự.

4.2 Phân tích sự biến đổi giá trị NDVI

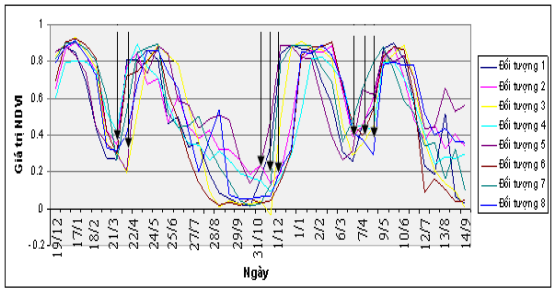
Thông thường, nếu chỉ số NDVI đạt giá trị cao (từ 0,5 - 0,9) là những vùng có thực vật phát triển tốt lúa đang ở giai đoạn đẻ nhánh, đòng hoặc trổ, hoặc vùng cây công nghiệp, cây ăn trái, rừng. Nếu giá trị < 0,5 thì là vùng không có hoặc có thực vật nhưng phát triển kém như vùng chuyên tôm, làm muối, ngập nước hay lúa mới sạ, lúa bắt đầu đâm chồi. Đối tượng không canh tác theo mùa vụ thì chỉ số ổn định suốt năm. Các đối tượng có giá trị NDVI dao động không cao theo thời gian sẽ được phân tách thành một đối tượng riêng. Đối với vùng canh tác lúa ở ĐBSCL, biểu đồ biến động NDVI theo dạng hình sin, giá trị đạt cực đại vào khoảng 0,8-1,0 tương ứng với giai đoạn cây phát triển tốt nhất và giảm xuống vào khoảng 0–0,4 khi kết thúc mùa vụ, sau đó giá trị này lại tiếp tục gia tăng theo quy luật trên khi bắt đầu một vụ mùa mới.



Hình 3: Ảnh chỉ số NDVI 15/10/2008

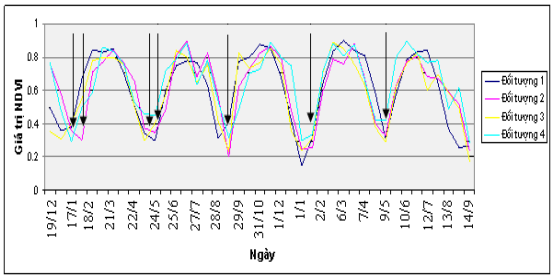
4.3 Thời gian xuống giống

Dựa vào biểu đồ biến động theo thời gian của giá trị NDVI ở những vùng trồng lúa để xác định được thời điểm xuống giống của từng mùa vụ Hình 4, Hình 5.



Ghi chú: ↓ *thể hiện khoảng thời gian xuống giống*

Hình 4: Sự biến động theo thời gian của những vùng lúa 2 vụ (ĐX chính vụ-HT sớm)



Hình 5: Sự biến động theo thời gian của những vùng lúa 3 vụ (ĐX muộn–HT–TĐ)

Kết quả phân tích cho thấy thời gian xuống giống cụ thể của từng mùa vụ lúa ở ĐBSCL như sau:

Vụ Đông Xuân: Thời gian xuống giống tập trung từ 15/10 đến 22/10; 24/11 đến 16/1 năm sau ở hầu hết các tỉnh. Ngày xuống giống tập trung cao nhất vào các đợt: Đợt 1: Từ 15/10 đến 22/10 tập trung ở huyện Trần Văn Thời (Cà Mau), Phước Long (Bạc Liêu), Mỹ Xuyên (Sóc Trăng), An Biên, Châu Thành, Cà Mau (Kiên Giang) và rãi rác ở một số tỉnh. Đợt 2: Từ 24/11 đến 25/12 tại Kiên Giang, Cần Thơ, Đồng Tháp, Vĩnh Long, Hậu Giang, Tiền Giang, Trà Vinh. Đợt 3: Từ 1/1 đến 16/1 tại An Giang, Sóc Trăng, Cầu Kè, Tiểu Cần, Trà Cú, Châu Thành (Trà Vinh), Ba Tri (Bến Tre), Đức Hòa, Tân An, Châu Thành, Chợ Gạo, Gò Công Đông, Gò Công Tây (Long An. Đợt 4: Từ 25/1 đến 1/2 tại các huyện Vĩnh Hưng, Mộc Hóa, Tân Thạnh (Long An).

Vụ Xuân Hè: Xuống giống vào nửa đầu tháng ba, tập trung chủ yếu từ 5/3 - 12/3 ở Cần Thơ, Hậu Giang, các huyện Sa Đéc, Thạnh Hưng, Châu Thành, Tháp Mười tỉnh Đồng Tháp, huyện Cai Lậy tỉnh Tiền Giang.

Vụ Hè Thu: Năm 2008 diện tích xuống giống cao trong khoảng thời gian từ 5/3 đến 16/6. Tuy nhiên, không phải diện tích xuống giống của cả đồng bằng đều cao trong khoảng thời gian này mà nó chỉ tập trung vào một số ngày cụ thể. Đợt 1: Từ 21/3–28/3 rãi rác ở Cần Thơ, Đồng Tháp và một phần ở An Giang, Kiên Giang, Trà Vinh. Đợt 2: Tập trung xuống giống với diện tích cao từ 6/4-13/4 tại các tỉnh Kiên Giang, An Giang, Đồng Tháp, và các huyện Vĩnh Hưng, Mộc Hóa, Tân Thạnh tỉnh Long An. Đợt 3: Từ 8/5-15/5 thời gian này xuống giống tương đối tập trung ở các tỉnh Long An, Bến Tre, Trà Vinh, Sóc Trăng, Bạc Liêu, Cà Mau và Kiên Giang. Và từ 24/5-31/5 là vụ Hè Thu muộn chỉ xuống giống ở vùng đã canh tác vụ Xuân Hè.

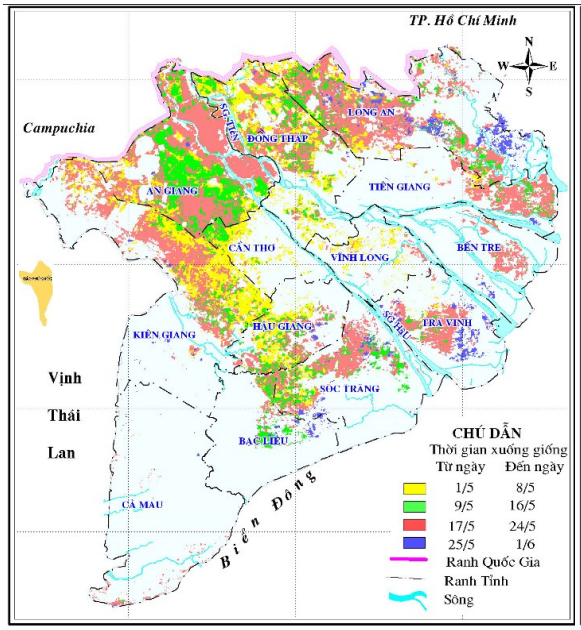
Vụ Thu Đông: Thời điểm xuống giống năm 2008 tương đối tập trung hơn năm 2009. Năm 2008 từ ngày 9 đến 16 tháng 6 và 27/7 đến 20/9 trong đó từ ngày 4 đến 11 tháng 8 diện tích xuống giống tương đối thấp; Năm 2009 xuống giống rãi rác không phân biệt rõ với vụ Hè Thu tập trung vào các ngày 2/6 – 9/6; 18/6 – 25/6; 4/7 - 11/7; 20/7 – 27/7 ; 15/9 - 1/10.

Vụ Mùa: Thời gian cấy vào khoảng từ tháng 8 đến giữa tháng 9 (Mạ được gieo sạ từ tháng 6, tháng 7). Đây là vụ lúa tương đối phức tạp vì có nơi xuống giống (sạ, cấy) bằng giống lúa mùa địa phương; hoặc giống trung mùa, hoặc giống ngắn ngày (90 - 100 ngày) gọi là mùa cao sản được thống kê vào diện tích lúa Thu Đông.

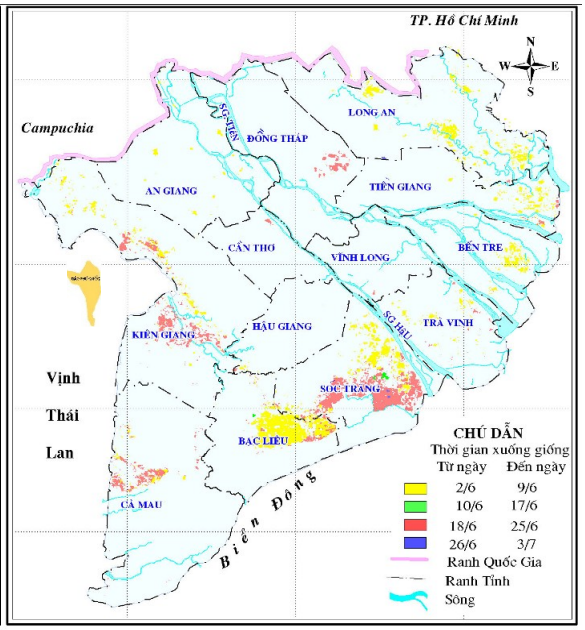
Diện tích cụ thể được thống kê theo từng tỉnh tháng 5, 6 năm 2009 được liệt kê Bảng 3, Hình 6, Hình 7.

Bảng 3: Kết quả giải đoán diện tích xuống giống lúa (ha) ở các tỉnh tháng 5, 6 năm 2009

Từ	Đến	An Giang	Bạc Liêu	Bến Tre	Cà Màu	Cần Thơ	Đồng Tháp	Hậu Giang
01/05	08/05	30.712	850	87	56	21.412	67.468	26.106
09/05	16/05	78.262	12.456	618	237	7.750	29.912	14.187
17/05	24/05	166.600	10.600	18.156	5.518	5.050	74.500	29.343
25/05	01/06	650	3.231	1.331	550	6,250	406	1.450
02/06	09/06	2.356	35.856	6.550	2.537	37	37	400
10/06	17/06		156					
18/06	25/06		4.131	68	9.431	400	4.012	168
26/06	03/07							
Từ	**Đến**	**S Trăng**	**Tiền Giang**	**Trà Vinh**	**K Giang**	**Long An**	**V Long**	**Tổng**
01/05	08/05	7.093	2.543	1.468	77.843	20.506	10.531	14.543
09/05	16/05	20.075	5.643	3.512	15.050	26.431	50	9.206
17/05	24/05	52.768	45.418	38.337	105.862	135.131	5.543	8.930
25/05	01/06	6.318	1.006	11.375	1.662	23.737	31	12.412
02/06	09/06	17.400	6.500	4.375	9.243	17.287	6	10.881
10/06	17/06	868						
18/06	25/06	39.868	1.106	943	18.625	225	175	2.225
26/06	03/07	243	206			25		206

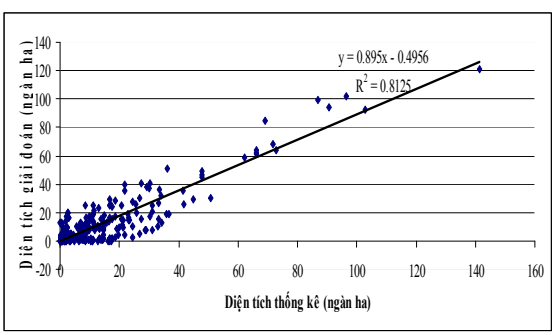


Hình 6: Bản đồ thời gian xuống giống trong tháng 5 năm 2009



Hình 7: Bản đồ thời gian xuống giống trong tháng 6 năm 2009

Đánh giá độ tin cậy: Kết quả giải đoán đã được thống kê so sánh với các số liệu báo cáo của địa phương thấy kết quả giải đoán có độ tin cậy khá cao (R^2=0,8125) có thể đề nghị được sử dụng trong theo dõi hiện trạng các trà lúa và cơ cấu mùa vụ (Hình 8).



Hình 8: Sự tương quan giữa diện tích mạ theo số liệu thống kê và số liệu giải đoán

4.4 Hiện trạng trà lúa và cảnh báo dịch hại trên lúa giải đoán từ ảnh MODIS NDVI

Khái niệm trà lúa có thể hiểu là vùng canh tác lúa mà tại đó cây lúa có cùng thời gian sinh trưởng và đồng nhất về thời gian gieo sạ ở một quy mô nhất định. Hiện trạng trà lúa phản ánh được tình hình xuống giống và có thể dự báo được khả năng dịch bệnh trên ruộng lúa theo từng trà lúa. Bản đồ trà lúa tỉnh An Giang cũng được xây dựng định kỳ 8 ngày/lần tương ứng với chu kỳ chụp ảnh của vệ tinh MODIS. Mỗi bản đồ được xây dựng với 5 đơn vị bản đồ tương ứng 5 giai đoạn sinh trưởng của lúa (Bảng 4).

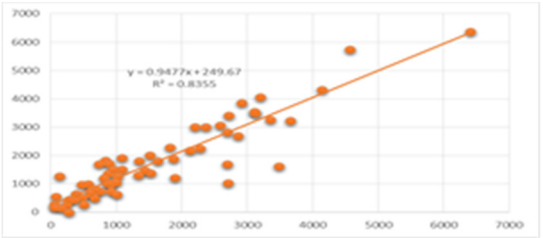
Bảng 4: Các giai đoạn sinh trưởng của lúa tương ứng với số ngày sau khi sạ (NSKS)

STT	Giai đoạn	Số ngày tuổi
1	Mạ	00-20 NSKS
2	Đẻ nhánh	21-40 NSKS
3	Đòng trổ	41-70 NSKS
4	Trổ chín	71-90 NSKS
5	Thu hoạch	>90 NSKS

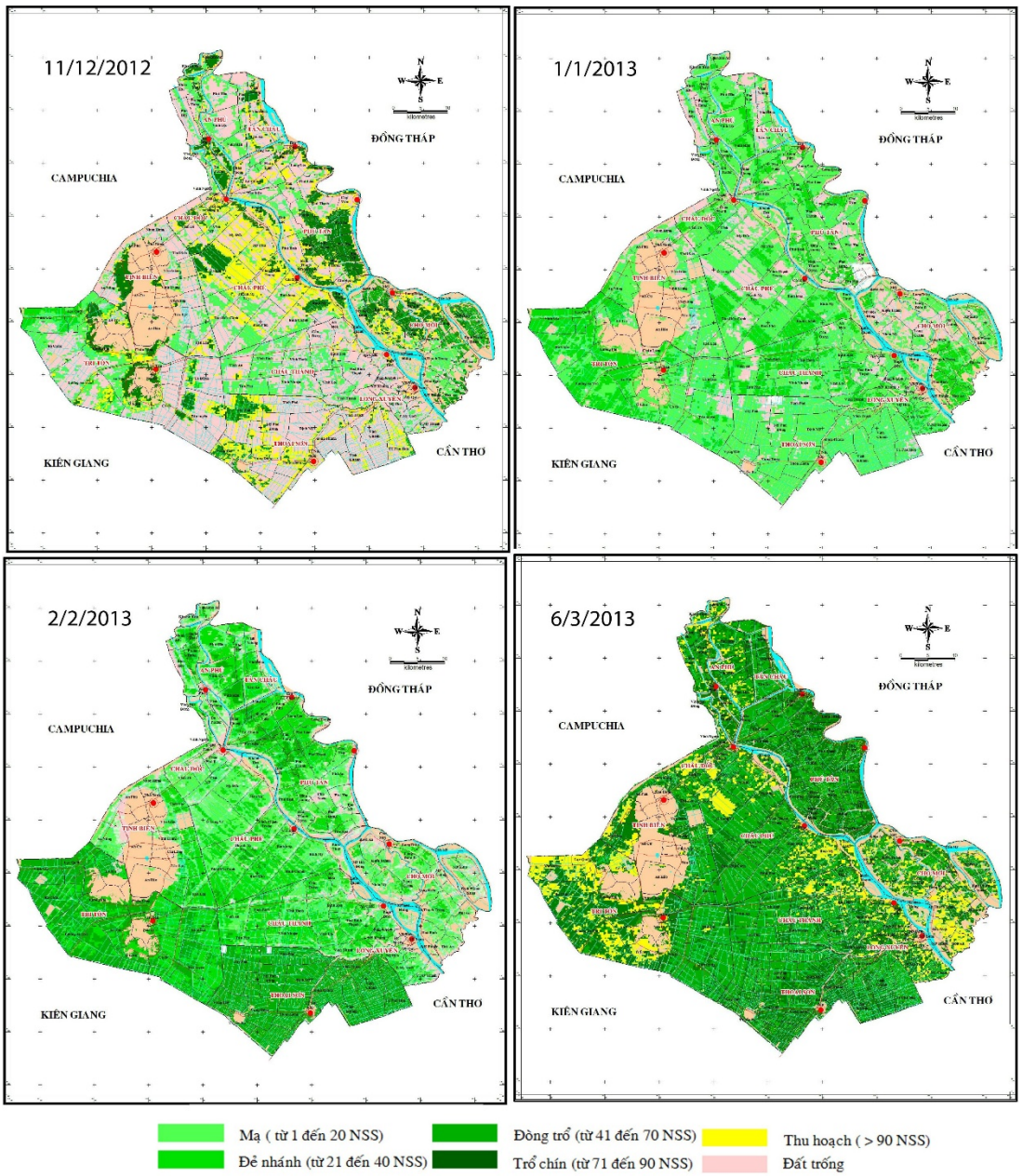
Hiện trạng các trà lúa được giải đoán xác định dựa trên các thời điểm xuống giống vụ lúa. Các bản đồ kết quả trà lúa bên dưới được phân tách từ ảnh MODIS NDVI theo từng thời điểm quan sát. Các trà lúa được xây dựng dựa trên chỉ số NDVI, tiến độ xuống giống của vụ, các số liệu điều tra thống kê và kết quả đi khảo sát thực địa, định vị GPS. Kết quả phân tách được 5 trà lúa, bao gồm: mạ, đẻ nhánh, đòng trổ, trổ chín và thu hoạch. Nhận thấy các trà lúa luôn hiện diện ở từng thời điểm quan sát. Các trà phụ thuộc vào tình hình, tiến độ và kế hoạch thực hiện mùa vụ trên địa bàn.

Các trà lúa sau giải đoán được tính toán độ tin cậy, kết quả cho thấy: Diện tích giải đoán và số liệu thống kê có mối tương quan chặt với nhau với hệ số R^2=0,8 (Hình 9) đồng thời hệ số tương quan của các trà lúa tương đối cao với R = 0,88 ở giai đoạn trổ chín; R = 0,82 ở giai đoạn đòng trổ; 0,72 ở

giai đoạn đẻ nhánh. Tuy nhiên, giai đoạn mạ và thu hoạch có tương quan thấp với hệ số R lần lược là 0,53 và 0,47 (Bảng 3). Nguyên nhân là do chưa phân tách được cụ thể diện tích xuống khi giải đoán, có sự nhầm lẫn khi phân loại trên ảnh NDVI,… Do đó, kết quả cảnh báo dịch hại có độ tin cậy cao đối với trà lúa ở giai đoạn đoạn đòng trổ và chín, kế đến là giai đoạn đẻ nhánh. Tương quan diện tích các trà lúa giải đoán với số liệu điều tra khác nhau theo giai đoạn sinh tưởng của các trà lúa và mùa vụ canh tác.



Hình 9: Mối tương quan giữa diện tích giải đoán với dữ liệu thống kê vụ Đông Xuân 2012-2013



Hình 10: Kết quả phân tách trà lúa từ ảnh MODIS vụ Đông Xuân 2012-2013

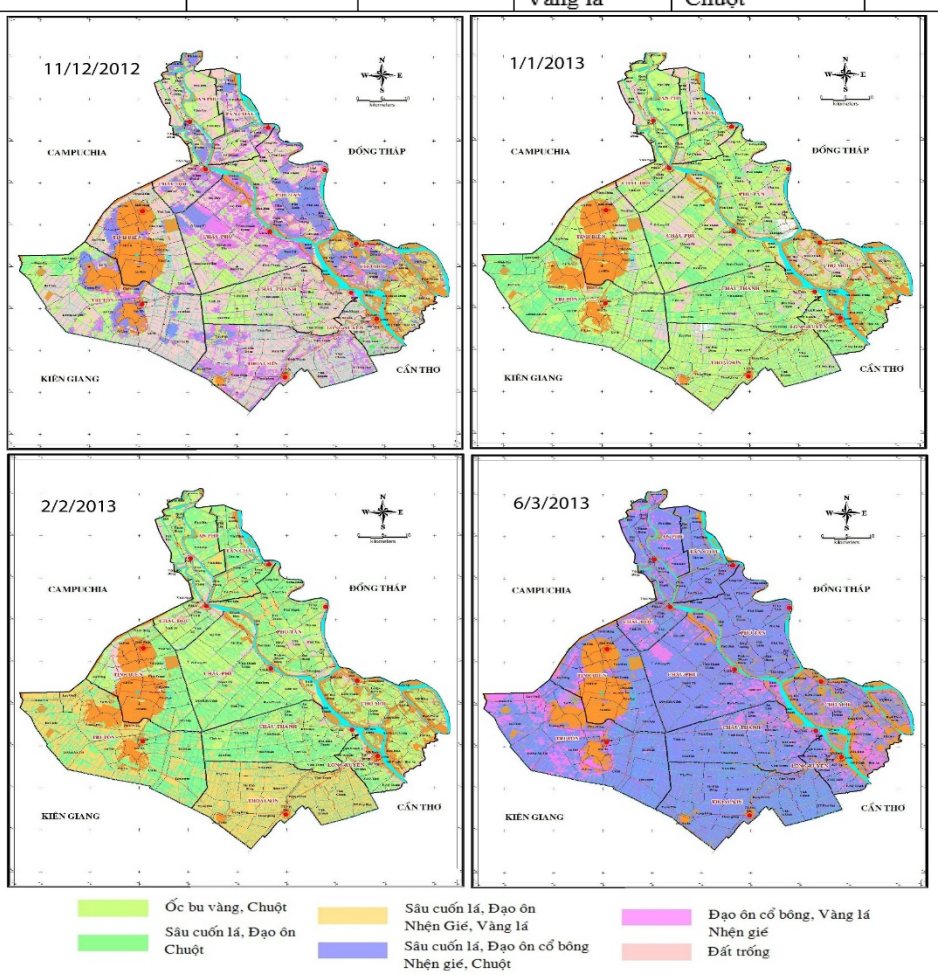
Bảng 3: Tương quan diện tích các trà lúa giải đoán so với số liệu điều tra

Diện tích theo kết quả giải đoán (ha)	Diện tích theo số liệu điều tra (ha)				
	Mạ	**Đẻ nhánh**	**Đòng-Trổ**	**Chín**	**Thu hoạch**
Mạ	**0,53**	0,00	-0,43	-0,29	-0,15
Đẻ nhánh	0,53	**0,72**	-0,10	-0,35	-0,19
Đòng-Trổ	-0,29	0,24	**0,82**	0,02	-0,10
Chín	-0,31	-0,28	0,37	**0,88**	0,46
Thu hoạch	-0,29	-0,33	0,06	0,63	**0,47**

Hệ thống cảnh báo dịch hại được xây dựng dựa trên loạt bản đồ hiện trạng trà lúa và các loại dịch hại gây hại thường xuyên trên các trà lúa đó. Do vậy, tương ứng với một trà lúa sẽ có vài loại dịch hại chủ đạo có khả năng xuất hiện cao (Bảng 4). Kết quả dự báo dịch hại lúa được trình bày ở Hình 11.

Bảng 4: Các loại dịch hại chính theo từng trà lúa (Nguồn: Chi cục BVTV AG, 2012)

	Giai đoạn sinh trưởng				
	Mạ	**Đẻ nhánh**	**Đòng trổ**	**Trổ chín**	**Thu hoạch**
Dịch hại gây hại thường xuyên	Ốc Bưu Vàng	Sâu cuốn lá	Sâu cuốn lá	Sâu cuốn lá	Đạo ôn cổ bông
	Chuột	Đạo ôn	Đạo ôn	Đạo ôn cổ bông	Vàng lá
		Chuột	Nhện gié	Nhện gié	Nhện gié
			Vàng lá	Chuột	



Hình 11: Hệ thống các bản đồ cảnh báo dịch hại được xây dựng dựa trên trà lúa trong vụ Đông Xuân 2012-2013 tỉnh An Giang

Từ bản đồ phân tách trà lúa (Hình 10) và bản đồ cảnh báo dịch hại (Hình 11) cho thấy vụ Đông Xuân năm 2012 - 2013 các loại dịch hại được cảnh báo vào một số ngày như sau:

Ngày 11/12/2012 trên địa bàn tỉnh An Giang phần lớn là đất trống, một số nơi ở các huyện Phú Tân, Chợ Mới, xã Ô Long Vĩ huyện Chợ Mới được cảnh báo các loại dịch hại sâu cuốn lá, đạo ôn cổ bông, nhện gié, chuột do các trà lúa ở giai đoạn trổ chín. Các trà lúa ở các xã khu vực phía Bắc huyện Châu Phú đang ở giai đoạn thu hoạch do đó các loại dịch hại như: Đạo ôn cổ bông, vàng lá, nhện gié được cảnh báo.

Ngày 1/1/2013 diện tích đất trống vẫn còn ở một vài xã của huyện Chợ Mới, Châu Phú còn lại hầu hết các trà lúa trên địa bàn tỉnh An Giang đã xuống giống và đang ở giai đoạn mạ, đây là điều kiện tốt cho ốc bưu vàng và chuột tấn công. Một số trà lúa ở huyện Tri Tôn đã chuyển sang giai đoạn đẻ nhánh do đó các loại dịch hại như sâu cuốn lá, đạo ôn, chuột được cảnh báo.

Ngày 2/2/2013 các loài dịch hại sâu cuốn lá, đạo ôn, nhện gié, vàng lá được cảnh báo tập trung ở các trà lúa đang ở giai đoạn đồng trổ của huyện Thoại Sơn, Tri Tôn. Phần lớn các trà lúa ở các huyện An Phú, Tân Châu, Tịnh Biên, Châu Thành đang ở giai đoạn mạ và một số đã chuyển sang giai đoạn đẻ nhánh do đó dịch ốc bưu vàng và chuột được cảnh báo cho các trà lúa ở giai đoạn mạ; dịch sâu cuốn lá, đạo ôn, chuột được cảnh báo cho các trà lúa ở giai đoạn đẻ nhánh.

Ngày 6/3/2013 đa số các trà lúa trên địa bàn tỉnh An Giang ở giai đoạn trổ chín và loại dịch hại sâu cuốn lá, đạo ôn cổ bông, nhện gié và chuột được cảnh báo trên toàn khu vực.

5 KẾT LUẬN

Nghiên cứu đã xây dựng hoàn chỉnh phương pháp sử dụng ảnh viễn thám MODIS trong xây dựng bản đồ theo dõi tiến độ xuống giống lúa phục vụ quản lý thời vụ và bản đồ cảnh báo dịch hại định kỳ 8 ngày/lần phục vụ công tác bảo vệ thực vật trên cây lúa.

Kết quả đánh giá độ tin cậy của bản đồ thông qua việc đối chiếu với dữ liệu khảo sát thực tế và dữ liệu báo cáo thống kê từ các cộng tác viên địa phương cho thấy có sự tương quan khá cao giữa kết quả giải đoán với dữ liệu báo cáo thực địa.

Sử dụng công nghệ viễn thám và GIS trong dõi thời gian xuống giống bằng cách tính toán sự biến động của chỉ số thực vật NDVI đa thời gian là rất hiệu quả, khách quan, phản ánh trung thực diễn biến về hiện trạng các trà lúa và cơ cấu mùa vụ mà không phụ thuộc vào tính chủ quan của con người. Từ kết quả này có thể xây dựng được hệ thống cảnh báo dịch lúa. Với sự kết hợp của công nghệ viễn thám và GIS là công cụ hỗ trợ công tác bảo vệ thức vật một cách hiệu quả, tốn ít chi phí, nhân lực và thời gian, có thể cung cấp thông tin cho các nhà quản lý có biện pháp xử lý kịp thời.

TÀI LIỆU THAM KHẢO

1. Aronoff, S. (1989) Geographic Information Systerm: A Management Perspective. WDL, Ottawa, Canada.
2. Burrough, P.A. (1986) Principles of Geographical Information Systems for Land Resources Assessement. Clarendon, Oxford.
3. Casadesús, J., Tambussi, E., Royo, C. and Araus, J. L. (2000), "Growth assessment of individual plants by an adapted remote sensing technique", *Durum Wheat Improvement in the Mediterranean Region: New challenges*, pp.129-132. Options Mediterranénnes, Zaragoza.
4. Lê Sâm (1996), Thủy *văn công trình*, NXB Khoa học kỹ thuật, Hà Nội.
5. Parida, B. R., B. Oinam, N. R. Patel, N. Sharma, R. Kandwal and M. K. Hazarika. (2008). Land surface temperature variation in relation to vegetation type using MODIS satellite data in Gujarat state of India. International Journal of Remote Sensing,Vol. 29, pp. 4219–4235.
6. Pham Van Du, Le Cam Loan. (2007) Review : Improvement of the rice breeding in intensive cropping system in the Mekong delta. Omonrice 15: 12-20 (2007). Cuu Long Delta Rice Research Institute, Can Tho, Vietnam
7. Trần Thị Hiền, Võ Quang Minh, (2010). Ảnh viễn thám Modis trong theo dõi tiến độ xuống giống trên vùng đất trồng lúa ở ĐBSCL. Kỷ yếu Hội thảo Ứng dụng GIS toàn quốc 2010. Nhà xuất bản Nông nghiệp. Số 198-2010/CXB/209-05/NN. Trang 85-93.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

MỘT KHUNG THỨC TÌM KIẾM VÀ TÁI SỬ DỤNG HÀM API TỰ ĐỘNG DỰA TRÊN ĐẶC TẢ HÌNH THỨC

Huỳnh Tấn Khải, Bùi Hoài Thắng, Quản Thành Thơ và Nguyễn Minh Thông[1]
[1] *Khoa Khoa học & Kỹ thuật Máy tính, Trường Đại học Bách Khoa Thành phố Hồ Chí Minh*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
A framework for automatic search and reusage of API functions based on formal specification

Từ khóa:
Tìm kiếm hàm tự động, gọi hàm tự động, tái sử dụng thành phần, đặc tả hình thức, thiết kế thoả thuận

Keywords:
Automatic API functions searching, automatic function composition, component reuse, formal specification, design by contract

ABSTRACT

To help programmers quickly develop their software systems, research communities have proposed some approaches for automatic search of function APIs, mostly based on keywords used in the API descriptions. However, the accuracy of those approaches is still limited. Formal specification with the rigor of the syntax and semantics can be applied to improve the research results of this direction. In this paper, we firstly present a survey on formal specification techniques for software design. Next, we propose a framework for searching and reusing API functions automatically based on formal specification described by JML. In addition, the framework also supports automatic composition of API functions to meet the requirement demand. Some case studies are also presented to imply the practical potential of our approach.

TÓM TẮT

Để giúp cho các lập trình viên phát triển nhanh hệ thống phần mềm của họ, cộng đồng nghiên cứu đã đề xuất một số cách tiếp cận cho việc tìm kiếm tự động các hàm API. Hầu hết các nghiên cứu này đều dựa trên các từ khóa được sử dụng trong các mô tả của các hàm API. Tuy nhiên, độ chính xác của các hướng tiếp cận này vẫn còn hạn chế. Đặc tả hình thức với sự chặt chẽ trong cú pháp và ngữ nghĩa có thể được áp dụng để cải thiện kết quả của các hướng nghiên cứu này. Trong bài báo này, đầu tiên, chúng tôi trình bày một khảo sát về các kỹ thuật đặc tả hình thức cho thiết kế phần mềm. Tiếp theo, chúng tôi đề xuất một khung thức cho việc tìm kiếm và tái sử dụng các hàm API tự động dựa trên đặc tả chính thức được mô tả bởi JML. Ngoài ra, khung thức còn hỗ trợ việc tổ hợp tự động các hàm API để đáp ứng các yêu cầu phần mềm. Một số kết quả thử nghiệm cũng được trình bày để cho thấy khả năng thực tế trong hướng tiếp cận của chúng tôi.

1 GIỚI THIỆU

Quá trình thiết kế nói chung trong hầu hết các ngành kỹ thuật đều có sự tái sử dụng các *thành phần* (component) đã được xây dựng sẵn. Trong công nghệ phần mềm, việc tái sử dụng các thành phần phần mềm là điều đã được đề cập và áp dụng từ rất sớm [1]. Việc xây dựng phần mềm dựa trên cơ sở tái sử dụng các thành phần đã có sẽ làm cho chi phí và thời gian xây dựng thấp hơn và chất lượng cũng tăng lên [2]. Tuy nhiên, bên cạnh các lợi điểm, việc tái sử dụng các thành phần hiện tại cũng gặp phải một số hạn chế trong việc ứng dụng thực tế. Trong đó khó khăn thường gặp nhất là việc lựa chọn thành phần tái sử dụng. Do số lượng các

thành phần được xây dựng sẵn trong các thư viện là rất lớn, các lập trình viên thường phải tốn khá nhiều công sức để tìm hiểu và lựa chọn thành phần phù hợp với nhu cầu của mình.

Trong lĩnh vực tái sử dụng thành phần thì việc tìm kiếm và gọi thực thi các hàm thư viện API (Application Programming Interface) là công việc được thực hiện thường xuyên nhất. Đã có nhiều nghiên cứu nhằm hỗ trợ việc tìm kiếm hàm thư viện đáp ứng yêu cầu của người lập trình. Tuy nhiên, hầu hết các hướng nghiên cứu này chủ yếu dựa trên việc so trùng chuỗi (từ khóa), nghĩa là dựa vào các mô tả bằng ngôn ngữ tự nhiên của hàm thư viện và của lời yêu cầu gọi hàm. Chẳng hạn như [3] nghiên cứu về máy tìm kiếm hàm dựa trên truy vấn ngôn ngữ tự nhiên; Assieme [4] là máy tìm kiếm mã nguồn dựa trên từ khóa; SNIFF [5] thực hiện các truy vấn bằng ngôn ngữ tự nhiên và trả về các đoạn trích mã nguồn phù hợp với mục đích của câu truy vấn. Với các cách tiếp cận trên thì chúng ta nhận thấy một điều là độ chính xác của các công cụ tìm kiếm này rất khác nhau, thậm chí với một công cụ tìm kiếm, nhưng trong các tình huống khác nhau cũng sẽ cho mức độ đúng đắn khác nhau. Mức độ chính xác của các công cụ trên có sự khác nhau là do chúng làm việc trên ngôn ngữ tự nhiên, vốn luôn chứa đựng sự nhập nhằng và đa nghĩa.

Một nhược điểm khác của các nghiên cứu này là chúng chỉ có thể tìm ra một hàm cụ thể đáp ứng yêu cầu (phù hợp theo từ khóa truyền vào). Khả năng kết hợp nhiều hàm lại với nhau để tổng hợp thành một lời giải hiện vẫn đang gặp nhiều khó khăn, vì để làm được điều này đòi hỏi chương trình tổng hợp phải "hiểu" được ngữ nghĩa của các hàm thư viện, thay vì chỉ có thể nhận dạng được các từ khoá mô tả các hàm thư viện này.

Một trong những hướng tiếp cận có thể vượt qua các khó khăn trên là sử dụng các *đặc tả hình thức* (formal specification) [6] để tìm kiếm và tổng hợp các hàm thư viện. Với cách tiếp cận này, các đặc tả hình thức, thực chất là các mệnh đề logic, với sự hỗ trợ của các công cụ suy luận và chứng minh định lý có thể cho được một lời giải đúng và đầy đủ cho một yêu cầu tìm kiếm hàm thư viện. Bên cạnh đó, việc sử dụng ngôn ngữ đặc tả hình thức sẽ loại bỏ được tính nhập nhằng và đa nghĩa của ngôn ngữ tự nhiên. Đồng thời, hướng tiếp cận này có thể hỗ trợ việc đưa ra một lời giải tổng hợp bao gồm một tập các hàm lồng ghép với nhau để có thể đáp ứng được yêu cầu, điều mà các phương pháp tìm kiếm dựa trên từ khóa là không thể làm được.

Trong các phương pháp phát triển phần mềm dựa trên đặc tả hình thức hiện nay thì phương pháp *Design by contract* (DBC) [7] nổi lên như một trong những phương pháp phát triển nhanh và chặt chẽ. Ý tưởng chính của phương pháp này là một lớp hoặc một phương thức và khách hàng của nó (đối tượng sử dụng) có một "*thỏa thuận*" (contract) với nhau. Thỏa thuận này thể hiện thông qua *tiền điều kiện* (pre-condition) và *hậu điều kiện* (post-condition). Khách hàng phải đảm bảo các điều kiện trước khi gọi một phương thức và kết quả trả về từ phương thức cũng phải thỏa được các yêu cầu mà khách hàng cần. Trong bài báo này, chúng tôi sẽ đưa ra một khung thức dựa trên hướng tiếp cận "thoả thuận" để tìm kiếm và tổng hợp các hàm thư viện nhằm đáp ứng một yêu cầu lập trình ban đầu.

Các đóng góp của chúng tôi bao gồm:

– Chúng tôi đưa ra một khảo sát chi tiết về các hướng tiếp cận đặc tả hình thức theo phương pháp "thoả thuận".

– Chúng tôi đưa ra một khung thức để tự động tìm kiếm và tổng hợp các hàm API dựa trên các đặc tả hình thức theo phương pháp "thoả thuận". Kết quả trả về có thể là một hàm đáp ứng yêu cầu của người gọi, hoặc cũng có thể là một chuỗi các hàm gọi tuần tự hoặc lồng nhau.

– Trong khung thức đề nghị, chúng tôi đưa ra hướng tiếp cận tính toán độ tương tự giữa các đặc tả hình thức của các API. Nhờ đó, khung thức cho phép sử dụng các kỹ thuật xử lý thông minh của lĩnh vực trí tuệ nhân tạo như *lập kế hoạch* (planning) [8] để tìm kiếm và tổng hợp các API cần thiết một cách hiệu quả.

Phần còn lại của bài báo này được tổ chức như sau: Phần 2 trình bày về các ngôn ngữ đặc tả hình thức theo phương pháp "thỏa thuận". Phần 3 trình bày ý tưởng về khung thức cho việc tìm kiếm hàm tự động dựa trên đặc tả hình thức JML. Phần 4 trình bày về một số kết quả thí nghiệm. Cuối cùng, Phần 5 là kết luận của bài báo.

2 CÁC NGÔN NGỮ HÌNH THỨC ĐẶC TẢ THEO PHƯƠNG PHÁP THỎA THUẬN

2.1 Đặc tả hình thức và hướng tiếp cận thoả thuận

Đặc tả hình thức cho phần mềm đã được quan tâm từ lâu trong lĩnh vực khoa học máy tính. Từ cuối những năm 1940, Turing đã cho rằng các chương trình tuần tự có thể được tạo ra một cách đơn giản bằng cách đánh chú thích các trạng thái của các thuộc tính của hệ thống tại các thời điểm

cụ thể [6]. Vào cuối những năm 1960, Floyd, Hoare và Naur đề xuất kỹ thuật tiền đề để chứng minh sự nhất quán giữa các chương trình tuần tự và các mô tả của chúng, được gọi là các đặc tả [9][10][11]. Sự quan tâm về các đặc tả hình thức và các ứng dụng đa dạng của chúng trong lĩnh vực công nghệ phần mềm đã phát triển liên tục cho đến hiện nay [12][13].

Một trong những kết quả mới và có tính thực tiễn nhất trong hướng tiếp cận này là việc đưa ra các "thoả thuận" giữa yêu cầu của một hàm và hiện thực của hàm đó [7]. Như đã giới thiệu trong Phần 1, một "thoả thuận" được biểu diễn thông qua các tiền điều kiện và hậu điều kiện. Ví dụ, một "thỏa thuận" cho phương thức *sqrt* (nhận vào một số và trả về căn bậc hai của nó) được biểu diễn bằng ngôn ngữ JML (*Java Modeling Language*) [14] như mô tả trong Hình 1. Trong ví dụ này, mệnh đề *require* thể hiện tiền điều kiện, là biến *x* truyền vào phải lớn hơn hoặc bằng *0*. Mệnh đề *ensure* thể hiện hậu điều kiện. Ở ví dụ này, giá trị trả về (*\result*) phải thỏa điều kiện là bình phương của nó phải bằng lại giá trị của biến *x* truyền vào.

```
//@ requires x >= 0.0;
//@ ensures \result * \result = x
public static double sqrt(double
x)
{ /*...*/ }
```

Hình 1: "Thỏa thuận" của phương thức *sqrt* được biểu diễn bằng ngôn ngữ JML

2.2 Các ngôn ngữ đặc tả hình thức theo hướng tiếp cận "thoả thuận"

2.2.1 Ngôn ngữ JML

JML (Java Modeling Language) [14] là ngôn ngữ đặc tả hành vi, được sử dụng để đặc tả hành vi của các mô đun trong ngôn ngữ Java. Nó là sự kết hợp phương pháp "thỏa thuận" với phương pháp đặc tả dựa trên mô hình của ngôn ngữ Larch [15]. JML sử dụng logic Hoare [16] kết hợp với việc đặc tả các tiền/ hậu điều kiện (pre-/post-condition) và các bất biến (invariant) của các thuộc tính, các cấu trúc.

Các đặc tả JML được thêm vào mã Java dưới dạng các chú giải (annotation) bên trong các chú thích (comment). Các chú thích trong ngôn ngữ Java được hiểu như là các chú giải JML khi nó bắt đầu bằng ký tự "@". Ví dụ trong Hình 2 minh họa việc sử dụng JML đặc tả cho lớp *Purse*. Trong đó, trường *balance* được đặc tả với bất biến (***invariant***) là *balance* luôn trong đoạn từ 0 đến *MAX_BALANCE*. Trường *pin* là một mảng thuộc kiểu ***byte*** với bất biến là *pin* có đúng 4 phần tử và các phần tử chỉ có giá trị từ 0 đến 9.

```
public class Purse {
  final int MAX_BALANCE;
  int balance;
  //@ invariant 0 <= balance &&
            balance <= MAX_BALANCE;

  byte[] pin;
  /*@ invariant pin != null &&
        pin.length == 4 &&
       (\forall int i; 0<=i&&i<4;
         0<=pin[i]&&pin[i]<=9);
*/

  /*@ requires amount >= 0;
   @ assignable balance;
   @ ensures
      balance==\old(balance)-
amount
       && \result == balance;
   @ signals (PurseException)
      balance == \old(balance);
*/
   int debit(int amount) throws

PurseException;
   }
```

Hình 2: Ví dụ minh họa về đặc tả JML

Phương thức *debit* nhận vào đối số *amount*. Phương thức này sẽ trừ *amount* vào *balance* và trả về giá trị của *balance* sau khi trừ trong trường hợp *balance* lớn hơn hoặc bằng *amount*. Trong trường hợp *balance* nhỏ hơn *amount*, một ngoại lệ *PurseException* sẽ được trả về. Tiền điều kiện (***requires***) của phương thức này là *amount* phải lớn hơn hoặc bằng 0; hậu điều kiện (***ensures***) là *balance* bị trừ đi một lượng *amount* và trả về kết quả sau khi trừ. Phương thức này còn có một hậu điều kiện nữa cho trường hợp phương thức trả về ngoại lệ (***signals***).

2.2.2 Spec#

Spec# [17] là một ngôn ngữ hình thức cho các "thỏa thuận" API (có sự ảnh hưởng từ JML), nó là sự mở rộng của ngôn ngữ C# với việc bổ sung các tiền/ hậu điều kiện và bất biến cho các đối tượng. Spec# là ngôn ngữ đặc tả cho phép kiểm tra động cũng như kiểm tra tĩnh chương trình. Hệ thống Spec# cung cấp một kiến trúc hoàn chỉnh, bao gồm trình biên dịch, thư viện hỗ trợ và các công cụ phát triển. Các đặc tả của Spec# cũng được biên dịch thành một phần của chương trình thực thi, trong đó

chúng sẽ được kiểm tra tự động. Đi kèm với Spec# là một bộ kiểm tra chương trình tự động, có tên là Boogie [18]. Nó cho phép kiểm tra các đặc tả theo kỹ thuật phân tích tĩnh [19].

Hình 3 minh họa về đặc tả của Spec# cho lớp *ArrayList*. Trong đó, phương thức *Insert* được đặc tả với các tiền điều kiện (***requires***), hậu điều kiện (***ensures***). Đặc biệt, Spec# còn cho phép các đặc tả có thể thừa kế lẫn nhau. Điều này giúp cho các đặc tả của Spec# rõ ràng, đáng tin cậy và dễ đọc hơn so với các ngôn ngữ đặc tả khác. Chi tiết về cách thức sử dụng Spec# chúng ta có thể tham khảo trong [17].

```
class ArrayList {
  public void Insert
        (int index, object
value)
  requires 0 <= index &&
                 index <= Count
;
  requires !IsReadOnly &&

!IsFixedSize;
  ensures Count == old(Count)+1;
  ensures value == this[index];
  ensures foral{int i in
0:index;

old(this[i])==this[i]};
  ensures foral{int i in
              index:old(Count);
       old(this[i])==this[i
+1]};
  { /*...*/ }
  //… …
}
```

Hình 3: Minh họa về đặc tả Spec#

2.2.3 Frama-C và ngôn ngữ ACSL

Frama-C [20] là một nền tảng dành riêng cho việc phân tích tĩnh mã nguồn chương trình được viết bằng ngôn ngữ C. Frama-C tập hợp một số kỹ thuật phân tích tĩnh [19] vào trong một khung thức duy nhất. Để được phân tích bởi Frama-C, các chương trình C cần được đặc tả bởi ngôn ngữ ACSL (ANSI/ISO C Specification Language) [21]. Đây là ngôn ngữ đặc tả hình thức cho phép chúng ta đặc tả các thuộc tính của một chương trình C. ACSL cũng được "lấy cảm hứng" từ ngôn ngữ JML.

Khái niệm quan trọng nhất trong ACSL đó là *thỏa thuận* hàm. Một thỏa thuận hàm cho một hàm *f* trong ngôn ngữ C là một tập các yêu cầu trên các đối số của *f* và một tập các thuộc tính cần được đảm bảo sau khi kết thúc việc thực hiện *f*. Chúng ta hãy xem xét ví dụ về hàm *max* (Hình 4). Hàm này không cần tiền điều kiện, không cần bất kỳ sự ràng buộc nào trên các đối số truyền vào. Hậu điều kiện của chúng mô tả rằng kết quả trả về của hàm *max* luôn lớn hơn hoặc bằng giá trị của *x*, *y* và sẽ bằng một trong hai tham số *x* hoặc *y* truyền vào.

```
/*@ ensures \result >= x &&
                      \result >=
y;
    ensures \result == x ||
                   \result == y;
*/
int max (int x, int y)
{ return (x > y) ? x : y; }
```

Hình 4: Minh họa về đặc tả ACSL

Ngoài tiền và hậu điều kiện, ACSL còn hỗ trợ đặc tả bất biến cho các đối tượng và vòng lặp. Tuy nhiên, ACSL lại không có cơ chế xử lý ngoại lệ như JML. Chi tiết về ACSL chúng ta có thể tham khảo trong [21].

2.2.4 Jahob

Jahob [22] là một hệ thống kiểm tra dành cho các chương trình viết bằng ngôn ngữ Java (tập con của ngôn ngữ Java). Sử dụng Jahob, các nhà phát triển có thể chứng minh theo phương pháp phân tích tĩnh rằng các phương thức hiện thực có phù hợp với các "thỏa thuận" đặc tả của chúng hay không. Jahob còn cho phép đặc tả các bất biến về các cấu trúc dữ liệu quan trọng cũng như các ràng buộc thiết kế. Ý tưởng cơ bản của Jahob là mô hình hóa các trạng thái của chương trình và các cấu trúc dữ liệu mà nó sử dụng thành các tập trừu tượng các đối tượng và các mối quan hệ giữa các đối tượng. Jahob sẽ hỗ trợ phát biểu các ràng buộc trên tập các đối tượng này. Chúng ta xét ví dụ về đặc tả của Jahob cho lớp *List* như được mô tả trong Hình 5.

Trong ví dụ trên, đặc tả của lớp *List* sử dụng biến đặc tả *content* để chỉ tập các đối tượng thể hiện trong danh sách. Tập này không tồn tại khi chương trình chạy – nó đơn giản là một sự trừu tượng. Chương trình Jahob chỉ sử dụng nó cho mục đích đặc tả và bộ kiểm tra Jahob sử dụng nó để kiểm tra chương trình. Chi tiết về cấu trúc và cú pháp của ngôn ngữ đặc tả cho hệ thống Jahob chúng ta có thể tham khảo trong [22].

```
class List
{
  /*: public static specvar
            content :: objset;
*/

  public List()
  /*: modifies content
      ensures "content = {}"  */
  public void add(Object o)
  /*: requires "o ~: content &
                       o ~=
null"
      modifies content
      ensures "content =
          old content Un {o}"
*/

  public void remove (Object o)
  /*: requires "o : content"
      modifies content
      ensures "content =
           old content - {o}"
*/
}
```

Hình 5: Minh họa đặc tả cho lớp *List* theo hệ thống Jahob

2.2.5 ADA 2012

Ada [23] là ngôn ngữ lập trình do Bộ quốc phòng Mỹ đầu tư phát triển vào khoảng nửa đầu thập niên 80 của thế kỷ 20. Qua nhiều năm phát triển, phiên bản Ada 2012 [24] có rất nhiều thay đổi, trong đó có một số bổ sung đáng kể như sau: hỗ trợ đặc tả hình thức trong chương trình, cho phép khai báo các tiền và hậu điều kiện cho các thủ tục, các bất biến của kiểu dữ liệu. Tất cả các sự bổ sung này nhằm mục đích kiểm tra tính đúng của chương trình cần xây dựng.

Hình 6 mô tả một ví dụ về đặc tả cho gói *Stack* bằng Ada 2012. Trong ví dụ này, biến *Stack* được khai báo với bất biến kiểu (*Type_Invariant*) là các phần tử trong Stack không trùng nhau. Các hàm được đặc tả ngay sau phần khai báo và bắt đầu bằng từ khóa *with*. Tiền điều kiện được khai báo với từ khóa *Pre* theo cú pháp *Pre* => <*expression*>. Tương tự, hậu điều kiện được khai báo với từ khóa *Post*.

```
package Stacks is
  type Stack is private;
    with Type_Invariant =>

Is_Unduplicated(Stack);
  function Is_Empty(S: Stack)
                  return
Boolean;
  function Is_Full(S: Stack)
                  return
Boolean;
  function Is_Unduplicated
       (S: Stack) return
Boolean;
  procedure Push(S: in out
Stack;
                 X: in Item)
    with
      Pre => not Is_Full(S),
      Post => not Is_Empty(S);
//...
end Stacks;
```

Hình 6: Minh họa đặc tả cho gói Stack bằng ngôn ngữ Ada 2012

2.3 So sánh các ngôn ngữ đặc tả

Bảng 1 tổng kết và so sánh lại các ngôn ngữ đặc tả dựa trên một số các tiêu chí. Trong số các ngôn ngữ đặc tả theo phương pháp "thỏa thuận" nêu trên thì cách sử dụng, cách thức làm việc của các phương pháp là khá giống nhau.

Hầu hết các ngôn ngữ trên đều bắt đầu từ ý tưởng nền tảng của JML, đây là ngôn ngữ tiên phong cho các ngôn ngữ đặc tả theo phương pháp "thỏa thuận". JML có ưu điểm là cấu trúc và cú pháp rõ ràng, đặc tả dựa trên ngôn ngữ chủ là ngôn ngữ Java đã trở nên rất phổ biến. Sự tách biệt giữa mã đặc tả và mã xử lý giúp cho hai công việc này hầu như tách biệt nhau. JML cho phép lồng ghép các đặc tả phi hình thức chung với các đặc tả hình thức nhằm hỗ trợ nhiều mức đặc tả cho chương trình.

Ngoài các ưu điểm của bản thân ngôn ngữ, JML còn có một số ưu điểm khác mà chúng tôi quyết định sử dụng nó cho khung thức của mình. Đó là: thư viện API của ngôn ngữ Java đã được đặc

tả đầy đủ bằng ngôn ngữ JML; JML có nhiều công cụ hỗ trợ cho việc kiểm tra chương trình như ESC/Java [25], KeY [26],... Do vậy, trong phần tiếp theo của bài báo này, JML là ngôn ngữ đặc tả hình thức được chúng tôi lựa chọn để tiếp tục phát triển cho khung thức tìm kiếm, tổng hợp và tái sử dụng hàm API tự động được đề cập trong các phần sau.

Bảng 1: So sánh các ngôn ngữ đặc tả hình thức theo hướng thoả thuận

	JML	Spec#	ACSL	Jahob	ADA 2012
Ngôn ngữ chủ	Java	Mở rộng của C#	C	Tập con của Java	ADA 2012
Cách thức đặc tả	Chèn vào dưới dạng chú giải trong các chú thích	Đưa thêm từ khóa vào ngay trong ngôn ngữ	Chèn vào dưới dạng chú giải trong các chú thích	Chèn vào dưới dạng chú giải trong các chú thích	Đưa thêm từ khóa vào ngay trong ngôn ngữ
Các thành phần đặc tả	Tiền điều kiện Hậu điều kiện Bất biến Ngoại lệ	Tiền điều kiện Hậu điều kiện Bất biến Ngoại lệ	Tiền điều kiện Hậu điều kiện Bất biến Hỗ trợ con trỏ	Tiền điều kiện Hậu điều kiện Bất biến Cấu trúc dữ liệu	Tiền điều kiện Hậu điều kiện Bất biến
Mục đích chính	Kiểm tra tính đúng của phương thức được hiện thực so với đặc tả	Kiểm tra tính đúng của phương thức được hiện thực so với đặc tả	Kiểm tra tính đúng của hàm được hiện thực so với đặc tả	Kiểm tra tính đúng của cấu trúc dữ liệu được hiện thực so với đặc tả	Kiểm tra tính đúng của hàm/ thủ tục được hiện thực so với đặc tả
Hỗ trợ đặc tả phi hình thức	Có	Không có	Không có	Không có	Không có
Số lượng công cụ hỗ trợ/ sử dụng	Nhiều (ECS/Java, KeY, JMLUnint,…)	Boogie	Frama-C	Hệ thống Jahob	Bản thân trình biên dịch ADA 2012

3 VÀ TÁI SỬ DỤNG HÀM API TỰ ĐỘNG

Hình 7 mô tả khung thức cho việc tìm kiếm và tái sử dụng hàm API tự động. Khung thức này có ba khối chức năng chính sau đây:

3.1 Bộ đánh giá sự tương tự giữa các đặc tả

Khối chức năng này giải quyết bài toán cơ bản để thực hiện việc tìm kiếm các đặc tả phù hợp với một yêu cầu đầu vào. Đó là đánh giá sự tương tự giữa các công thức logic đặc tả các hàm. Nhờ vậy, chúng ta có thể thu hẹp không gian tìm kiếm hàm. Việc thu hẹp này dựa trên sự tính toán mức độ tương tự của đặc tả yêu cầu với các đặc tả của hàm API có trong thư viện.

Nguyên lý cho việc phân loại hàm dựa trên học máy là với một tập các hàm đã cho (P) và một lời yêu cầu (c), giải thuật sẽ *tiên đoán* những hàm từ P có khả năng thỏa mãn được yêu cầu của c. Hướng tiếp cận này dựa trên một số định nghĩa sau:

Định nghĩa 1: Ma trận phụ thuộc

Gọi Γ là một tập đặc tả được biểu diễn dưới dạng logic của các hàm thư viện. Khi đó chúng ta định nghĩa:

$$\mu = \Gamma \times \Gamma \rightarrow \{0,1\}$$

$$\mu(c,p) = \begin{cases} 1 & \text{nếu } p \text{ có khả năng thỏa mãn được yêu cầu của } c \\ 0 & \text{ngược lại} \end{cases}$$

Định nghĩa 2: Ma trận đặc trưng

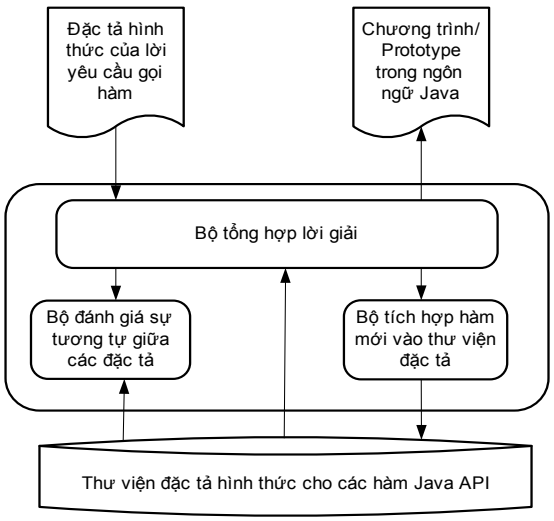
Gọi $T=\{t_1, ..., t_m\}$ là tập các mệnh đề (term) đặc tả trong Γ. Chúng ta định nghĩa hàm Φ như sau:

$$\Phi = \Gamma \times \{1, ..., m\} \rightarrow \{0,1\}$$

$$\Phi(c,i) = \begin{cases} 1 & \text{nếu } t_i \text{ xuất hiện trong } c \\ 0 & \text{ngược lại} \end{cases}$$

Định nghĩa 3: Hàm đặc trưng

Hàm đặc trưng φ: $\Gamma \rightarrow \{0,1\}^m$, với mỗi *c* thuộc Γ thì chúng ta có vector $\varphi^c \in \{0,1\}$ thỏa điều kiện $\varphi_i^c = 1 \Leftrightarrow \Phi(c,i) = 1$.



Hình 7: Khung thức cho việc tìm kiếm và tái sử dụng hàm API tự động

Định nghĩa 4: Hàm phân loại

Cho mỗi *p* trong Γ, hàm phân loại được định nghĩa như sau $C_p(.)$: $\Gamma \rightarrow R$, với mỗi lời yêu cầu *c* thì hàm phân loại $C_p(c)$ sẽ tính toán mức độ phù hợp của hàm *p* so với lời yêu cầu *c*.

Ví dụ, với các hàm *f*1, *f*2 và *f*3 được mô tả trong Hình 8, nếu chúng ta ký hiệu các mệnh đề p_1, p_2, p_3, p_4 tương ứng với các tập toán tử {+,-} {*}, {>, ≤}, {*log*}[1], chúng ta sẽ có ma trận đặc trưng như sau:

φ	p_1	p_2	p_3	p_4
f_1	0	1	0	0
f_2	1	0	0	0
f_3	1	0	1	1

```
//@ ensures \result == 2 * x;
double f1 (double x);

//@ ensures \result == - x;
double f2 (double x);

//@ ensures
//@    (x>0)==>(\result==x+1) &&
//@
(x≤0)==>(\result==log(x2));
double f3 (double x);
```

Hình 8: Một số đặc tả hàm bằng JML

[1] Đây là một sự phân loại toán tử thường gặp dựa trên độ ưu tiên,

Khi đó, chúng ta sẽ tính được vector đặc trưng cho từng lời gọi hàm dựa vào φ^c. Ví dụ dựa vào ma trận trên, chúng ta có vector đại diện cho lời gọi hàm f_2 là φ^{f_2} =[1,0,0,0] (giá trị 1 ở cột thứ *i* nghĩa là c_2 có đặc trưng *i*). Với vector φ vừa tính được, chúng ta có thể phát triển các phương pháp tính toán độ tương tự. Các phương pháp này có thể đơn giản là phép tích góc giữa hai vector, hoặc có thể là một phương pháp học máy như Naive Bayes [27], tuỳ theo độ phức tạp của bài toán.

3.2 Bộ tổng hợp lời giải

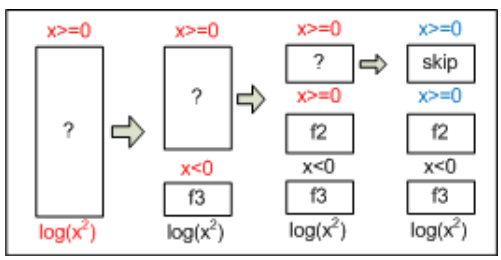
Khối chức năng thứ hai là bộ tổng hợp lời giải. Đây là bộ phận quan trọng nhất trong khung thức. Nhờ vào việc các hàm thư viện được đặc tả bằng các hàm logic và phương pháp tính độ tương tự giữa các đặc tả xây dựng được, chúng ta có thể phát triển một cơ chế tìm kiếm và tổng hợp các hàm API đáp ứng nhu cầu ban đầu dựa trên kỹ thuật *lập kế hoạch* trong lĩnh vực Trí tuệ Nhân tạo (AI Planning).

Bộ tìm kiếm hàm sử dụng kỹ thuật lập kế hoạch dựa trên trí tuệ nhân tạo hoạt động theo nguyên tắc sau: Xuất phát từ yêu cầu ban đầu (*c*) và tập các hàm *P* có độ tương tự trong ngưỡng quy định đối với (*c*), bộ lập kế hoạch sẽ chọn ra một hàm *p* từ *P* (theo độ ưu tiên về mức độ phù hợp), nếu tiền điều kiện của *p* phù hợp với tiền điều kiện của *c* thì quá trình tìm kiếm kết thúc và trả kết quả về. Ngược lại, giải thuật xem tiền điều kiện của *p* như là một hậu điều kiện mới và lặp lại quá trình tìm kiếm. Trong trường hợp quá trình tìm kiếm bế tắc, giải thuật sẽ quay lui lại với các hàm *p'* trong *P* ($p' \neq p$).

Với các hàm thư viện như mô tả trong Hình 8, giả sử chúng ta có một lời yêu cầu gọi hàm được đặc tả như sau:

```
//@ requires x ≥ 0
//@ ensures \result == log(x2)
```

Với thư viện hàm và lời yêu cầu như trên, quá trình *lập kế hoạch* của chúng ta được thực hiện thông qua các bước như trong Hình 9.



Hình 9: Quá trình tìm kiếm hàm dựa trên lập kế hoạch

Ban đầu, hệ thống sẽ cố gắng tìm một hàm có khả năng đưa ra kết quả là $log(x^2)$ từ yêu cầu ban đầu là $x \geq 0$. Sử dụng cơ chế tính toán tương tự giữa các đặc tả như mô tả trong Phần 3.1, bộ lập kế hoạch xác định được hàm phù hợp nhất là f_3. Để có thể tạo ra kết quả là $log(x^2)$, hàm f_3 yêu cầu ngõ nhập phải thoả điều kiện $x<0$. Như vậy bộ lập kế hoạch cần tìm một hàm thoả mãn *subgoal* : $\{x \geq 0 \Rightarrow x<0\}$. Một lần nữa, việc tính toán độ tương tự giữa các hàm sẽ xác định hàm phù hợp nhất cho subgoal này là f_2[2]. Đến đây, một lời giải tổng hợp đã được hình thành đầy đủ như sau: $f_3(f_2(x))$.

3.3 Bộ tích hợp hàm mới vào thư viện đặc tả

Chức năng này đóng vai trò giúp cho thư viện đặc tả các hàm của chúng ta phong phú thêm, giúp cho những lần sau khi gặp những yêu cầu tương tự, hệ thống sẽ có sẵn các hàm/ chuỗi các hàm đáp ứng yêu cầu, giúp tối ưu hiệu suất của hệ thống.

Kết quả đầu ra của khung thức này là một đoạn mã hoặc một nguyên mẫu phần mềm (prototype) trong đó có chứa các lời gọi hàm cần thiết theo yêu cầu ban đầu của người sử dụng. Ví dụ, sau khi tìm ra lời giải tổng hợp cho yêu cầu được mô tả trong Phần 3.2, đặc tả và lời giải này lại được tiếp tục lưu trữ vào thư viện như một hàm mới (hàm f_4) để phục vụ cho những yêu cầu tìm kiếm và tổng hợp lời giải mới sau này.

4 MỘT SỐ BÀI TOÁN MINH HOẠ

Sau đây là một số ví dụ minh họa việc sử dụng đặc tả JML để đặc tả cho các yêu cầu tìm kiếm và gọi hàm tự động từ một số trường hợp thực tế.

Trường hợp 1: Tìm ngay được một hàm phù hợp, đáp ứng yêu cầu.

Yêu cầu 1: "Cho một hình lập phương, cạnh có độ dài đại số là *a*. Tính thể tích của hình lập phương đó".

Chúng ta đặc tả yêu cầu trên như sau:

```
//@ requires a > 0
/*@ ensures \result > 0 &&
 @JMLDouble.approximatelyEqualTo
  @     (a^3, \result, 0.001);
 */
```

Bằng cách phân tích độ tương thích giữa hậu điều kiện (***ensure***) của yêu cầu với hậu điều kiện của các hàm có trong thư viện thì chúng ta nhận được một hàm phù hợp ngay, đó là hàm p*ow()* của lớp *java.lang.Math*. Hàm *pow()* này có đặc tả JML như sau[3]:

```
/*@ ensures
 @JMLDouble.approximatelyEqualTo
 @    (a^b, \result, 0.000001);
*/
public static double pow
            (double a, double b);
```

Tiền điều kiện của hàm *pow()* được bỏ qua nghĩa là tiền điều kiện luôn đúng. Đó đó, nó thỏa mãn với tiền điều kiện của yêu cầu. Như vậy *pow(a,3)* là hàm cần gọi.

Yêu cầu 2: "Chèn một phần tử *x* vào đầu Stack *S*".

Đặc tả JML cho yêu cầu:

```
//@ requires S.size<S.maxSize-1;
/*@ ensures
S.size==\old(S.size+1)
  @         &&
S.firstElement()==x;
 */
```

Với phương pháp tương tự như yêu cầu 1, khung thức cũng dễ dàng tìm được 1 hàm thỏa mãn yêu cầu trên là hàm *push()* của lớp Stack, với đặc tả như sau:

```
/*@ requires size < maxSize - 1;
 @ assignable size;
 @ ensures size = \old(size + 1)
          && S.firstElement()==x;
 @ ensures_redundantly (\forall
 @        int i; 0≤i && i<size-1;
 @element(i)==\old(element(i)));
 @*/
public void push(Object x);
```

Trường hợp 2: Cần tinh chế đặc tả mới có thể tìm ra lời gọi hàm.

Yêu cầu 3: "Trả về vị trí của phân tử *x* trong mảng *a*. Nếu phần tử *x* không tồn tại trong mảng *a* thì trả về -1".

Đặc tả JML cho yêu cầu trên như sau:

[2] Để tính toán được rằng đặc tả của f_2 ($x \Rightarrow -x$) là tương đương với đặc tả của subgoal, chúng ta sẽ cần đến một *công cụ chứng minh* (prover). Đã có rất nhiều prover được phát triển trong cộng đồng nghiên cứu, nhưng trong khuôn khổ của bài báo này, chúng tôi không thảo luận đến các prover.

[3] Đặc tả này đã có trong thư viện về đặc tả của các hàm thư viện của Java.

```
//@ requires a != null;
/*@ ensures \result==-1 ||
 @  (\result≥0 &&
\result<a.length
 @        && a[\result] == x);
*/
```

Căn cứ vào hậu điều kiện của đặc tả yêu cầu trên, chúng ta sẽ tìm được hàm *binarySearch()* của lớp *java.util.Arrays* với đặc tả như sau:

```
/*@ requires a != null &&
  @  (\forall int i; 0<i &&
  @    i<a.length; a[i-1]≤a[i]);
  @ ensures \result==-1 ||
  @
(\result≥0&&\result<a.length
  @      && a[\result]==key);
*/
public static int binarySearch
              (int[] a, int
key);
```

Tuy nhiên, tiền điều kiện của hàm *binarySearch()* lại chưa phù hợp với tiền điều kiện của yêu cầu. Do đó, quá trình tìm kiếm và tổng hợp lời giải lại tiếp tục như cơ chế đã được mô tả trong phần 3.2. Quá trình tìm này sẽ thu được hàm *sort()* của lớp *java.util.Arrays*. Hàm *sort()* có đặc tả như sau:

```
/*@ requires a != null;
  @ ensures  a != null &&
  @  (\forall int i; 0<i &&
      i<a.length; a[i-
1]≤a[i]);*/
public static void sort(int[]
a);
```

Đến đây, tiền điều kiện của hàm *sort()* này hoàn toàn phù hợp với tiền điều kiện của yêu cầu. Như vậy quá trình tìm kiếm sẽ kết thúc và chúng ta có một kế hoạch gọi thực thi các hàm như sau:

```
java.util.Arrays.sort(a);
java.util.Arrays.binarySearch(a,
x);
```

5 KẾT LUẬN

Việc áp dụng ngôn ngữ đặc tả hình thức vào cơ chế tìm kiếm và tái sử dụng thành phần tự động là một hướng nghiên cứu mới, giúp cho kết quả tìm kiếm sẽ tốt hơn so với các nghiên cứu dựa trên từ khóa. Cho đến nay, các nghiên cứu về việc ứng dụng đặc tả hình thức, đặc biệt là với hướng tiếp cận thoả thuận, chỉ chủ yếu tập trung vào việc kiểm chứng tính đúng của phần mềm. Trong bài báo này, chúng tôi đề xuất sử dụng các đặc tả hình thức vào việc tìm kiếm và tổng hợp các thành phần cần thiết theo nhu cầu của người lập trình. Từ đó, chúng tôi trình bày một khung thức để tìm kiếm và tổng hợp các API được mô tả bằng ngôn ngữ JML. Hơn thế nữa, cơ chế hoạt động của khung thức này cho phép lưu trữ và bổ sung các hàm thư viện mới từ kết quả tổng hợp được. Một số thí nghiệm ban đầu với những bài toán lập trình thường gặp đã thu được kết quả khả quan trên khung thức của chúng tôi. Điều này khích lệ chúng tôi công bố kết quả nghiên cứu của mình để thu thập thêm ý kiến đóng góp từ cộng đồng.

Hiện tại, khung thức do chúng tôi đề nghị chủ yếu dựa trên đặc tả JML và sử dụng để tìm kiếm và tổng hợp các hàm API cho ngôn ngữ Java. Tuy nhiên, do bản chất của hình thức đặc tả trong khung thức là các công thức logic Hoare [16] nên khung thức này hoàn toàn có thể tuỳ biến hiệu quả để áp dụng cho các ngôn ngữ lập trình dạng câu lệnh (imperative) khác như C, C++, C#, Scala, Pascal …

TÀI LIỆU THAM KHẢO

1. M. D. McIlroy, J. M. Buxton, P. Naur and B. Randell, 1968. Mass-produced software components. *In Proceedings of the 1st International Conference on Software Engineering*, pp: 88–98.
2. Sametinger, Johannes, 1997. *Software engineering with reusable components*. Springer, pp: 11-15.
3. Tran DinhHuy and HuaPhung Nguyen, 2013. API specification-based function search engine using natural language query. *Computing, Management and Telecommunications, International Conference on IEEE*, pp:140-145.
4. Hoffmann, Raphael, J. Fogarty, and D.S. Weld, 2007. Assieme: finding and leveraging implicit references in a web search interface for programmers. *Proc. of the 20th annual ACM symposium on User interface software*, pp:13-22.
5. Chatterjee, Shaunak, S. Juvekar, and K. Sen, 2009. Sniff: A search engine for java using free-form queries. In: *Fundamental Approaches to Software Engineering*. Springer Berlin Heidelberg, pp: 385-400.
6. B. Randell, 1973. *The Origin of Digital Computers*. Springer-Verlag, p. 298.

7. Meyer, Bertrand, 2002. *Design by Contract*. Prentice Hall.
8. D.S. Weld, 1999. Recent Advances in AI Planning. *AI Magazine*, Vol.20: 2-93.
9. R. Floyd, 1976. Assigning Meanings to Programs. *In Mathematical Aspects of Computer Science. Proc. Syrup. Appl. Maths*. Vol.19:19-32.
10. C.A.R. Hoare, 1969. An Axiomatic Basis for Computer Programming. In: *Comm. ACM*, Vol. 12: 576-583.
11. P. Naur, 1969. Proofs of algorithms by General Snapshots. BIT Numerical Mathematics. Vol.6.4: 310-316.
12. E.M. Clarke, J.M. Wing *et al*, 1996. Formal Methods: State of the Art and Future Directions. *ACM Computing Surveys*, Vol.28: 626-643.
13. J.M. Wing, J. Woodcock and J. Davies, 1999. FM-99: Worm Conference on Formal Methods in the Development of Computing Systems. *LNCS 1708 and 1709*, Springer-Verlag.
14. Burdy, Lilian, *et al*., 2005. An overview of JML tools and applications. *International Journal on Software Tools for Technology Transfer*. Vol.7: 212-232.
15. J.V. Guttag and J.J. Horning, 1993. *LARCH: Languages and Tools for Formal Specification*. Springer-Verlag.
16. Jacobs, Bart, and E. Poll, 2001. *A logic for the Java Modeling Language JML*. Springer Berlin.
17. Barnett, et.al, 2005. The Spec# programming system: An overview. *Construction and analysis of safe, secure, and interoperable smart devices*. Springer Berlin Heidelberg, pp:49-69
18. Barnett, Mike, *et al*., 2006. Boogie: A modular reusable verifier for object-oriented programs. *Formal methods for Components and Objects*. Springer Berlin Heidelberg. pp: 364-387.
19. Gopan, Denis, and Thomas, 2007. Guided static analysis. In: *Static Analysis*, Springer Berlin Heidelberg, pp: 349-365.
20. L. Correnson, *et al*., 2013. *Frama-C User Manual*, Software Safety Laboratory, Saclay.
21. P. Baudin, et.al, 2012. *ACSL: ANSI/ISO C Specification Language*. Version 1.6.
22. Kuncak, Viktor, and M. Rinard, 2006. An overview of the Jahob analysis system: project goals and current status. In: *Parallel and Distributed Processing Symposium*. 8 pp
23. Watt, A. David, B.A. Wichmann and W. Findlay, 1987. *Ada language and methodology*. Prentice Hall International (UK).
24. Barnes and John, 2011. Rationale for Ada 2012: Contracts and aspects. *ADA USER*, Vol.32 (4): 247-264.
25. C. Flanagan, *et.al*, 2002. Extended static checking for Java. *Proc. of the Conference on Programming Language Design and Implementation*, pp: 234-245.
26. Beckert, Bernhard, R. Hähnle, and P.H. Schmitt, 2007. *Verification of object-oriented software: The KeY approach*. Springer-Verlag.
27. Rish, Irina. An empirical study of the naive Bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol.3: 41-46.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

XÂY DỰNG CÔNG CỤ NGĂN CHẶN TRUY VIỆC CẬP WEB ĐEN (HÌNH ẢNH, NỘI DUNG)

Huỳnh Bé Thơ và Trương Quốc Định[1]
[1] *Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Building tool to prevent access to black web (image, content)

Từ khóa:
Lọc web, Máy học vector hỗ trợ (SVM), phân lớp văn bản, phân lớp hình ảnh

Keywords:
Web filtering, Support vector machine learning, text classification, image classification

ABSTRACT

Web filtering is used to prevent access to black web pages (web pages have desirable content or images). In this paper, we apply classification method with Support vector machine learning (SVM) to build a web filtering tool that is integrated with 2 filters: text filter – use text classification method and image filter – use image classification method. With two filters, this tool can prevent user access to desirable content web pages or remove desirable images when web page is displayed on the browser.

TÓM TẮT

Công cụ lọc web (web filtering) được sử dụng để ngăn chặn việc truy cập đến các trang web đen (trang web mang nội dung hoặc hình ảnh không mong muốn). Trong bài báo này, chúng tôi ứng dụng phương pháp phân lớp với Máy học vector hỗ trợ (SVM) để xây dựng một công cụ lọc web được tích hợp 2 bộ lọc: bộ lọc văn bản – phân lớp văn bản (text classification) và bộ lọc image – phân lớp hình ảnh (image classification). Với hai bộ lọc này, công cụ có thể cấm người dùng truy cập đến trang web có nội dung văn bản không mong muốn hoặc loại bỏ các hình ảnh không mong muốn khi hiển thị web lên trình duyệt.

1 GIỚI THIỆU

Cùng với sự phát triển của công nghệ ADSL và sự cạnh tranh giữa các nhà cung cấp dịch vụ dẫn đến Internet ngày càng phổ biến và dễ tiếp cận. Không chỉ có thể tiếp xúc với môi trường Internet tại cơ quan, trường học, quán net mà ngay tại gia đình, nhà trọ cũng thật dễ dàng sở hữu một đường truyền ADSL hay cáp quang tốc độ cao. Không thể phủ nhận Internet là kho tri thức khổng lồ, một công cụ đắc lực hỗ trợ cho việc học tập, nghiên cứu nhưng ngược lại nó cũng chứa nhiều mối nguy hiểm tiềm ẩn bên trong, cụ thể là sự lan tràn của các trang web chứa nội dung, hình ảnh không lành mạnh, không phù hợp với thuần phong mỹ tục của một số quốc gia...

Các chương trình lọc web trước đây thường sử dụng các công nghệ mang tính thủ công: ngăn chặn dựa trên từ khóa (keyword), chặn theo địa chỉ liên kết (URL hay IP),… Các công nghệ này đã lỗi thời bởi vì không phải bất cứ trang web nào có chứa các từ ngữ "nhạy cảm" đều là trang web khiêu dâm và cũng thật khó để ngăn chặn theo URL hoặc IP khi số lượng các trang web sex quá lớn và tăng thêm liên tục. Trong ngữ cảnh đó, xu hướng chung của các phần mềm lọc web đen ngày nay là dựa trên phân tích nội dung trang web sử dụng kỹ thuật khai mỏ dữ liệu (data mining).

Đi theo xu hướng chung đó, chúng tôi đã xây dựng một hệ thống lọc web dựa trên kỹ thuật phân lớp, cụ thể là phân lớp văn bản và hình ảnh với SVM.

Phần tiếp theo của bài báo được tổ chức như sau. Phần 2 trình bày các nghiên cứu liên quan và hướng tiếp cận của bài báo. Phần 3 giới thiệu các bước nghiên cứu bao gồm: bài toán phân lớp (phần 3.1), máy học Vector hỗ trợ SVM (phần 3.2), mô hình hóa văn bản (phần 3.3), cuối cùng là biểu diễn ảnh bằng đặc trưng SIFT và mô hình Bag of Words. Kết quả nghiên cứu trình bày ở phần 4. Kết luận và đề xuất được giới thiệu trong phần 5.

2 NGHIÊN CỨU LIÊN QUAN

2.1 Nghiên cứu liên quan

Các hệ thống lọc web dựa trên kỹ thuật phân lớp được phát triển gần đây thường dựa trên 3 hướng tiếp cận: phân lớp văn bản, phân lớp hình ảnh và phân lớp dựa trên sự kết hợp nhiều yếu tố (văn bản, hình ảnh/video). Theo xu hướng thứ nhất có thể kể đến nghiên cứu của Du *et al*.2003 [5], Kim *et al*.2006 [8] hoặc Santos *et al*.2012 [11]. Du và các đồng sự đã đề xuất một hệ thống lọc web sử dụng thuật toán tính độ tương đồng của vector văn bản với tập học để phân loại trang web khiêu dâm và không khiêu dâm. Một nghiên cứu khác của Kim phát triển hệ thống phân loại trang web đen thành nhiều cấp bậc bằng phương pháp phân lớp văn bản với máy học SVM. Nghiên cứu của Santos áp dụng giải thuật DMC (dynamic Markov compression) phân lớp văn bản để lọc các website khiêu dâm.

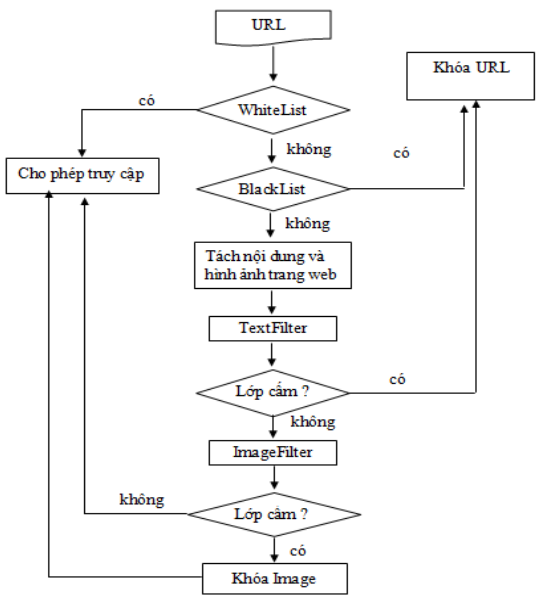
Phân loại trang web dựa theo kỹ thuật phân lớp hình ảnh cũng có nhiều nghiên cứu. Jiao *et al*.2011 [12] nghiên cứu ứng dụng kỹ thuật phân lớp hình ảnh sử dụng đặc trưng màu sắc và SVM cho hệ thống ngăn chặn các website khiêu dâm. Nghiên cứu của Zhao.2010 [13] đề xuất phương pháp kết hợp đặc trưng màu sắc, kết cấu và đặc trưng SIFT cho thuật toán phân lớp hình ảnh với SVM ứng dụng cho hệ thống phát hiện ảnh khỏa thân trên web.

Cuối cùng là phương pháp kết hợp thông tin văn bản và hình ảnh/videos để nhận diện web khiêu dâm. W.Hu *et al*. 2011 [6] đề xuất phân loại trang web dựa trên đặc trưng được kết hợp giữa văn bản, hình ảnh và video. Trang web được đại diện bởi vector $W = (w_1, w_2,..., w_t, w_{t+1}, w_{t+2}, ..., w_{t+7})$ với w_i $(1 \le i \le t)$ là trọng số các đặc trưng văn bản và w_{t+j} $(1 \le j \le 7)$ đại diện cho các đặc trưng thứ j của image/videos. Nghiên cứu của M. Hammami *et al*.2003 [7] thực hiện kết hợp đặc trưng văn bản và hình ảnh để xây dựng đặc trưng cho trang web, sau đó quyết định xem trang web này có thuộc web cấm hay không bằng giải thuật cây quyết định (decision tree). Saikat Sen. 2010 [10] giới thiệu một hệ thống lọc web sử dụng giải thuật phân lớp Naïve Bayes để phân loại trang web dựa vào các đặc trưng url, tiêu đề, từ khóa và nội dung (văn bản, hình ảnh) của trang web.

2.2 Hướng tiếp cận của bài báo

Tất cả các nghiên cứu trên đều thực hiện lọc web tiếng Anh và chỉ có chức năng lọc web (cho truy cập hay không), không có chức năng lọc ảnh khiêu dâm (vẫn truy cập được nhưng các ảnh khiêu dâm sẽ không hiển thị).

Trong bài báo này, chúng tôi giới thiệu một hệ thống lọc web dựa trên kỹ thuật phân lớp văn bản và hình ảnh với SVM. Tuy nhiên không thực hiện phân loại trang web bằng cách kết hợp 2 yếu tố nội dung văn bản và hình ảnh trang web mà là xây dựng 2 bộ lọc. Bộ lọc text (textfilter) thực hiện phân lớp văn bản tiếng Việt để phân loại trang web và bộ lọc image (imagefilter) dựa trên kỹ thuật phân lớp hình ảnh, thực hiện chức năng "lọc" hình ảnh trang web. Hình 1 thể hiện sơ đồ hoạt động của hệ thống



Hình 1: Sơ đồ hoạt động của hệ thống

3 CÁC BƯỚC NGHIÊN CỨU

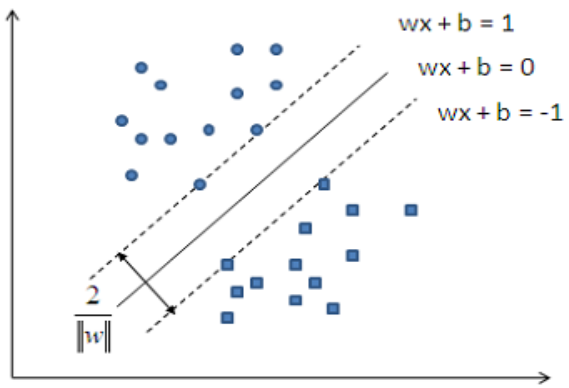
3.1 Bài toán phân lớp

Bài toán phân lớp (phân loại) là một bài toán kinh điển trong lĩnh vực khai mỏ dữ liệu. Mục tiêu của bài toán là xây dựng một mô hình phân lớp dựa trên tập dữ liệu học có nhãn (lớp) [14]. Ví dụ cho sẵn một tập dữ liệu các trang web được gán nhãn là web khiêu dâm hay bình thường, vấn đề là cần một

phương pháp huấn luyện để xây dựng một mô hình phân lớp từ tập dữ liệu mẫu này sau đó dùng mô hình này dự đoán lớp của những trang web mới (chưa biết nhãn).

3.2 Máy học Vector hỗ trợ - SVM

Phương pháp SVM [15] ra đời từ lý thuyết học thống kê do Vapnik và Chervonenkis xây dựng và có nhiều tiềm năng phát triển về mặt lý thuyết cũng như ứng dụng trong thực tiễn. SVM được đánh giá là 1 trong 10 giải thuật quan trọng của khai mỏ dữ liệu [14]. Các ứng dụng thực tế cho thấy, phương pháp SVM có khả năng phân loại khá tốt đối với bài toán phân loại văn bản cũng như trong nhiều ứng dụng khác (như nhận dạng chữ viết tay, phát hiện mặt người trong các ảnh, ước lượng hồi quy, ...).



Hình 2: Phân lớp tuyến tính với SVM

Bài toán cơ bản của SVM là bài toán phân loại hai lớp: Cho trước n điểm trong không gian d chiều (mỗi điểm thuộc vào một lớp kí hiệu là +1 hoặc –1, mục đích của giải thuật SVM là tìm một siêu phẳng (hyperplane) phân hoạch tối ưu cho phép chia các điểm này thành hai phần sao cho các điểm cùng một lớp nằm về một phía với siêu phẳng này. Hình 2 minh họa phân lớp với SVM trong mặt phẳng.

Xét tập dữ liệu mẫu có thể tách rời tuyến tính $\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$ với $x_i \in R^d$ và $y_i \in \{\pm 1\}$. Siêu phẳng tối ưu phân tập dữ liệu này thành hai lớp là siêu phẳng có thể tách rời dữ liệu thành hai lớp riêng biệt với lề (margin) lớn nhất. Tức là, cần tìm siêu phẳng H: y = w.x + b = 0 và hai siêu phẳng H1, H2 hỗ trợ song song với H và có cùng khoảng cách đến H. Với điều kiện không có phần tử nào của tập mẫu nằm giữa H_1 và H_2, khi đó:

w.x + b >= +1 với y = +1

w.x + b >= -1 với y = -1

Kết hợp hai điều kiện trên ta có:

y(w.x + b) >= 1.

Khoảng cách của siêu phẳng H_1 và H_2 đến H là ||w||. Ta cần tìm siêu phẳng H với lề lớn nhất, tức là giải bài toán tối ưu tìm $\min_{w,b}\|w\|$ với ràng buộc y(w.x + b) >= 1. Người ta có thể chuyển bài toán sang bài toán tương đương nhưng dễ giải hơn là $\min_{w,b}\frac{1}{2}\|w\|^2$ với ràng buộc y(w.x + b) >= 1. Lời giải cho bài toán tối ưu này là cực tiểu hóa hàm Lagrange:

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i(y_i(<w,x_i> +b) - 1)$$

Trong đó α là các hệ số Lagrange, α≥0. Sau đó người ta chuyển thành bài toán đối ngẫu là cực đại hóa hàm W(α):

$$\max_\alpha W(\alpha) = \max_\alpha (\min_{w,b} L(w,b,a)) \quad (2)$$

Từ đó giải để tìm được các giá trị tối ưu cho w,b và α. Về sau, việc phân loại một mẫu mới chỉ là việc kiểm tra hàm dấu sign(wx +b).

Giải thuật SVM cơ bản giải quyết được bài toán phân lớp tuyến tính, tuy nhiên nếu ta kết hợp SVM với phương pháp hàm nhân (kernel – based method), sẽ cho phép giải quyết một số bài toán phi tuyến bằng cách ánh xạ dữ liệu vào một không gian có số chiều lớn hơn. Không có bất kỳ thay đổi cần thiết nào về mặt giải thuật, việc duy nhất cần làm là thay thế các tích vô hướng của hai vector u.v bởi hàm nhân K(u,v). Chúng ta có một số hàm nhân cơ bản được dùng phổ biến được cho trong Bảng 1.

Bảng 1: Một số hàm nhân thường được dùng

Kiểu hàm	Công thức
Tuyến tính	K(u,v) = u.v
Đa thức bậc d	$K(u,v) = (u.v + c)^d$
Radial Basis Function	$K(u,v) = \exp(-\gamma\|u-v\|^2)$

Trong giới hạn nghiên cứu này, chúng tôi không đi sâu vào giải thuật SVM. Các mô hình phân loại trong nghiên cứu được thực hiện nhờ vào sự hỗ trợ của công cụ LibSVM [23].

3.3 Mô hình hóa văn bản

Để có thể thực hiện phân lớp văn bản với Máy học vector hỗ trợ, mỗi văn bản cần được biểu diễn dưới dạng vector với các thành phần (chiều) của vector này là các trọng số của các từ chỉ mục. Khái

niệm "từ chỉ mục" ở đây theo nghĩa là một chuỗi kí tự liên tiếp nhau trong văn bản, không nhất thiết phải là một từ có nghĩa trong ngôn ngữ [15]. Như vậy, giai đoạn đầu tiên trong việc vector hóa văn bản là thực hiện việc tách rời các từ. Tập hợp tất cả các từ để biểu diễn văn bản được rút ra từ tập các văn bản đang xét gọi là tập đặc trưng.

Việc tách các từ này một cách chính xác có ảnh hưởng rất lớn đến kết quả phân loại. Trong những năm gần đây, có nhiều công trình nghiên cứu về tách từ tiếng Việt. Tiêu biểu là công cụ vnTokenizer được phát triển trong đề tài VLSP của tác giả Lê Hồng Phương [22]. Công cụ này tách từ cho độ chính xác là 97%. Hình 3 minh họa một đoạn văn bản được tách từ bởi VnTokenizer.

Ngoài_ra /, / có / một / vấn_đề / nảy_sinh / trong / khi / tách / từ / là / việc / xuất_hiện / các /từ / mới /,/ đây / là / vấn_đề / không_thể / bỏ_qua / khi / ngôn_ngữ / là / luôn_luôn / thay_đổi / và / sinh_ra / các / từ / mới /./ Trong khi / từ_điển / (/ dành / cho /xử_lý / ngôn_ngữ_tự_nhiên /) / không_thể / cập_nhật /hết / được /./

Hình 3: Ví dụ tách từ với VnTokenizer

Rõ ràng rằng, các từ trong văn bản có mức độ quan trọng khác nhau đối với văn bản và cả đối với các văn bản khác trong tập văn bản cần phân loại. Một số từ như từ nối, dấu chấm câu, ký hiệu đặc biệt, từ chỉ số lượng ("và", "các", "những, "mỗi",…). không mang tính phân biệt trong khi phân loại. Để giảm bớt số lượng đặc trưng, nâng cao tốc độ tính toán, các từ gọi là stopword này cần được loại bỏ. Tuy nhiên các stopword này có số lượng không đáng kể, cần thiết phải áp dụng một giải thuật giúp chọn lựa các đặc trưng thật sự hữu ích cho việc phân lớp. Một vài thuật toán giúp lựa chọn các đặc trưng [9] như: ngưỡng tần suất văn bản (Document Frequency thresholding - DF), độ lợi thông tin (Information Gain - IG), thông tin tương hỗ (Mutual Information - MI), phương pháp thống kê χ^2 (CHI), độ mạnh của từ và một số phương pháp khác. Trong bài báo này, chúng tôi đã áp dụng phương pháp ngưỡng tần suất văn bản DF cho nghiên cứu này bởi tính đơn giản và hiệu quả của nó. Phương pháp này tính tần suất văn bản (DF) cho mỗi đặc trưng và loại bỏ những đặc trưng có tần suất văn bản nhỏ hơn ngưỡng cho trước. Các đặc trưng có tần suất văn bản thấp sẽ mang ít thông tin phân loại và thường là dữ liệu nhiễu.

Sau khi loại bỏ các stopword, các từ có tần số DF thấp, tập đặc trưng còn lại đó là tập hợp các từ "quan trọng" còn lại để biểu diễn văn bản. Việc phân loại văn bản sẽ dựa trên tập đặc trưng này.

Kế đến, mỗi văn bản trong tập đang xét sẽ được biểu diễn bởi các trọng số của từ. Cách tính trọng số của từ được áp dụng theo công thức trọng số TF*IDF (tf-idf weighting) [1]. Trọng số tf-idf được tính như sau:

$$TF\text{-}IDF_{t,d} = TF_{t,d} * \log(N/ DF_t) \quad (3)$$

– $TF_{t,d}$: là số lần xuất hiện của từ t trong văn bản d.

– DF_t : số lượng văn bản có chứa từ t.

– N: tổng số văn bản trong tập dữ liệu đang xét.

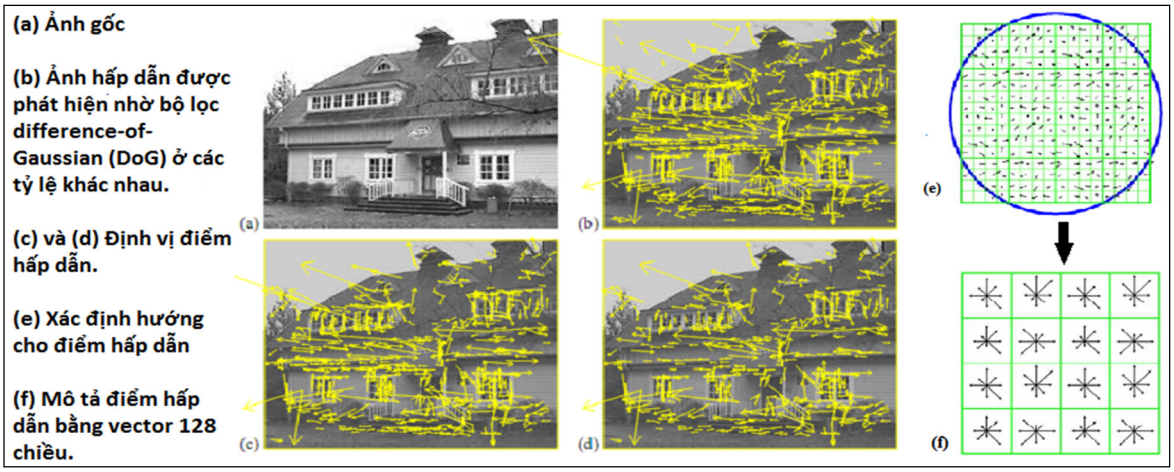
3.4 Biểu diễn ảnh bằng đặc trưng SIFT và mô hình BoW (Bag of Words)

Giống như một từ trong một văn bản text, một tấm ảnh cũng có thể xem là tập hợp các điểm hấp dẫn cục bộ hoặc các điểm nổi bật, là những vùng nhỏ (small regions) chứa nhiều thông tin cục bộ của ảnh – còn gọi là các đặc trưng. Có nhiều đặc trưng có thể được sử dụng để biểu diễn cho ảnh [16], trong đó có đặc trưng cục bộ. Người ta thường chia đặc trưng cục bộ thành 2 loại là những điểm trích xuất được từ điểm "nhô ra" (salient points) của ảnh và đặc trưng SIFT [2] được trích chọn từ các điểm hấp dẫn Haris (interest points). Điểm hấp dẫn này được mô tả bởi vector 128 chiều (gọi là các sift descriptor). Các vector này "bất biến" với việc thay đổi tỉ lệ ảnh, quay ảnh, đôi khi là thay đổi điểm nhìn và thêm nhiễu ảnh hay thay đổi cường độ chiếu sáng của ảnh. Hình 4 mô tả các bước rút trích đặc trưng SIFT.

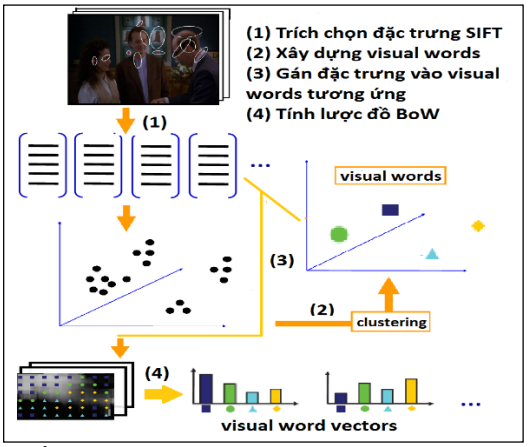
Bước kế tiếp, chúng tôi dùng mô hình bag of words (BoW) [3,4,20,21,25] để thực hiện kết tập các đặc trưng cục bộ SIFT. Mô hình này dùng một giải thuật (ví dụ như k-means) gom nhóm các đặc trưng cục bộ SIFT để xây dựng các visual words. Tập các visual words này gọi là codebook. Sau đó gán các đặc trưng cục bộ trên mỗi tấm ảnh vào visual words gần nhất. Khoảng cách Euclid thường được sử dụng để tính khoảng cách từ đặc trưng đến visual words gần nhất. Khoảng cách Euclid được tính bằng công thức sau:

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Ảnh sẽ biểu diễn ảnh bằng một vector với các thành phần là các visual words. Các vector này còn được gọi là các lược đồ BoW (BoW histograms). Trọng số của các visual words được tính theo **tần suất xuất hiện của visual word** trong mỗi ảnh. Hình 5 mô tả các bước tạo mô hình BoW để biểu diễn ảnh.



Hình 4: Các bước rút trích đặc trưng SIFT



Hình 5: Mô hình bag of words (BoW)

4 KẾT QUẢ VÀ THẢO LUẬN

Chúng tôi cài đặt một http proxy có tích hợp 2 bộ lọc TextFilter (lọc văn bản) và ImageFilter để thực hiện chức năng ngăn chặn việc truy cập vào các trang web đen. Khi người dùng truy cập vào trang web có nội dung cấm, công cụ sẽ chuyển trình duyệt về trang thông báo cấm truy cập hoặc sẽ hiển thị trang web nhưng "lọc" lại các ảnh khỏa thân (nếu có). Tùy vào số lượng ảnh có trên trang web, thời gian (được tính từ lúc bắt đầu nhập URL vào address bar đến lúc load xong toàn bộ trang web) để trình duyệt có tích hợp proxy với 2 bộ lọc load một trang web có thể chậm hơn vài giây so với trình duyệt thông thường.

4.1 Đánh giá bộ lọc văn bản - TextFilter

Tập dữ liệu cho bộ lọc TextFilter là 2518 mẫu tin được tải về từ internet. Các mẫu tin này thuộc 2 chủ đề (khiêu dâm và không khiêu dâm - thường). Chúng tôi dùng 1518 mẫu tin làm tập huấn luyện (dữ liệu học) và 1000 mẫu tin để kiểm tra (xem Bảng 2).

Bảng 2: Tập dữ liệu văn bản

STT	Tên tập	Khiêu dâm	Bình thường
1	Tập huấn luyện	900	618
2	Tập kiểm tra	500	500

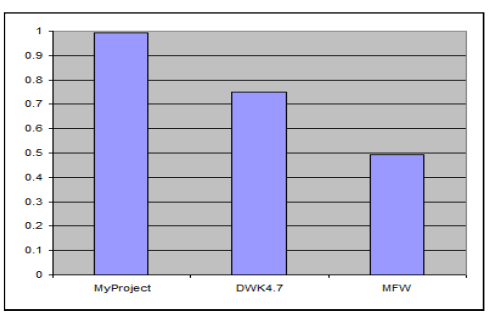
Tập huấn luyện được đặt trong 2 thư mục riêng biệt (0NonAdult: chứa mẫu tin bình thường và 1Adult: chứa mẫu tin khiêu dâm). Sử dụng thư viện VietTokenizer [22] để tách câu thành các "từ" riêng biệt. Tổng số đặc trưng thu được sau khi tách từ là hơn 50.000 đặc trưng. Thực hiện loại bỏ stopword và các đặc trưng có tần số DF<3, cho kết

quả còn lại là 19.587 đặc trưng. Biểu diễn các mẫu tin bằng vector trọng số TF*IDF. Nhãn của các mẫu tin sẽ được gán tự động theo thứ tự của thư mục chứa mẫu tin. Trong trường hợp này, các trang web thuộc nhóm bình thường có nhãn là 0 và các trang khiêu dâm có nhãn là 1.

Để phân lớp văn bản bằng SVM, chúng tôi sử dụng bộ thư viện LibSVM [23] với công cụ *grid.py* giúp lựa chọn các tham số tối ưu cho giải thuật SVM. Bằng nghi thức kiểm tra chéo (5-fold) trên tập học, *grid.py* đã tìm ra 2 tham số (C = 32 và gamma = 0.0001220703125) sau khi tính được kết quả phân loại cao nhất là 99,0119% trên tập huấn luyện.

Thực hiện kiểm nghiệm trên tập kiểm tra cho kết quả phân lớp đạt độ chính xác là 93,65%.

Để đánh giá hiệu quả thực tế của bộ lọc, chúng tôi thực hiện so sánh khả năng phát hiện các trang web khiêu dâm của công cụ với các phần mềm lọc web "thuần việt" hiện nay như: DWK 4.7 (Tác giả Vũ Lương Bằng – Công ty Điện Thoại Đông), MiniFireWall 4.0 - MFW (Tác giả Huỳnh Ngọc Ấn – Phòng Tin học, bưu điện Đồng Tháp). Các chương trình này ứng dụng kỹ thuật lọc theo URL hoặc keyword để chặn các trang web. Dữ liệu để so sánh là 122 mẫu tin lấy ngẫu nhiên từ 122 website khác nhau từ công cụ tìm kiếm Google với các từ khóa tìm kiếm là "truyện người lớn", "truyện khiêu dâm", "truyện sex", "truyện 18+". Kết quả so sánh thể hiện ở biểu đồ 1.



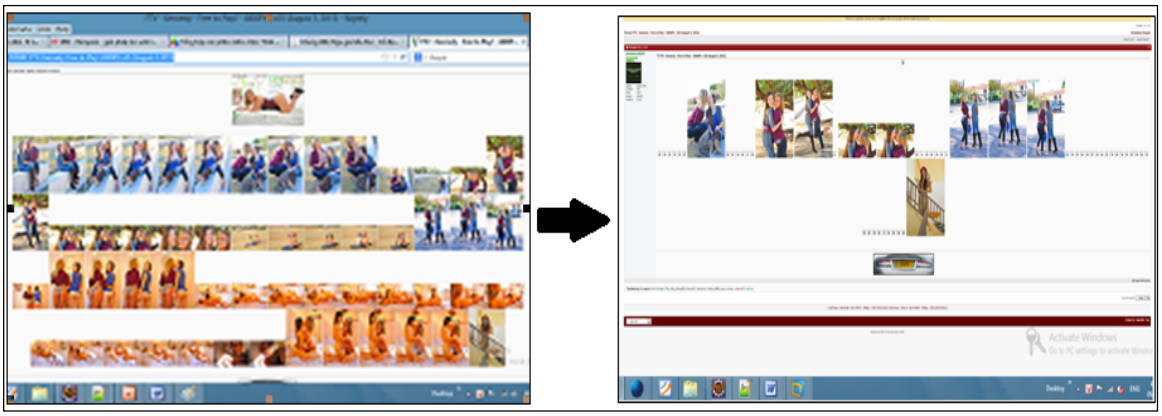
Biểu đồ 1: So sánh khả năng phát hiện website truyện khiêu dâm của công cụ (MyProject) với DWK và MFW

Kết quả này cho thấy việc sử dụng kỹ thuật phân lớp văn bản để phân loại, phát hiện và ngăn chặn các trang web đen thật sự hiệu quả hơn các kỹ thuật khác.

4.2 Đánh giá bộ lọc hình ảnh ImageFilter

Mô hình phân lớp cho bộ lọc ImageFilter được huấn luyện từ tập dữ liệu gồm 1905 ảnh (1066 ảnh thuộc nhóm ảnh khiêu dâm và 839 ảnh không khiêu dâm – bình thường). Tương tự như dữ liệu văn bản cho bộ lọc TextFilter, tập dữ liệu ảnh này cũng được đặt vào 2 thư mục khác nhau để dễ dàng trong việc gán nhãn. Để tìm vector đặc trưng SIFT của ảnh, chúng tôi sử dụng chương trình SiftDemoV4 [24] của David G. Lowe. Các vector này sẽ được gom nhóm bằng giải thuật k-means để xây dựng các visual words và thực hiện vector hóa tập dữ liệu ảnh với các visual words tìm được. Chúng tôi thí nghiệm với số lượng các visual words là 200, 300, 500, 1000, 2000, 3000 và N (Với N là căn bậc hai của tổng số đặc trưng). Ta thu được mô hình phân lớp với SVM cho độ chính xác ổn định nhất là (76.3255% và 80.1084%) ứng với số visual words là (200 và 3000). Từ thực nghiệm và nghiên cứu của [17] cho thấy quy mô của tập từ vựng có ảnh hưởng rất lớn đến tốc độ tính toán của hệ thống. Vì vậy với một ứng dụng chạy thời gian thực như trong nghiên cứu này, chúng tôi ưu tiên chọn mô hình với số visual words là 200, dù độ chính xác của mô hình này thấp hơn mô hình với số visual words 3000. Hình 6 cho ví dụ về kết quả phân lớp ảnh của bộ lọc ImageFilter. Ảnh bên trái là một trang web có chứa ảnh khiêu dâm và bên phải là trang web được hiển thị lên trình duyệt sau khi bộ lọc loại bỏ các ảnh khiêu dâm.

Trên thực tế các nghiên cứu về phát hiện ảnh khiêu dâm dựa vào đặc trưng SIFT có thể cho kết quả phân lớp rất khác nhau, tùy vào nhiều yếu tố như dữ liệu học, giải thuật phát hiện điểm đặc trưng cục bộ,... Nghiên cứu của Lopes *et al.*2009 [18] cho kết quả nhận diện ảnh khiêu dâm sử dụng đặc trưng SIFT với SVM là 65±3%. Nghiên cứu khác của Steel *et al.* [19], cho kết quả đánh giá dựa vào chỉ số TPR (true positive rates) là (0.58 và 0.66) tương ứng với FPR (false positive rates) tại (0.1 và 0.2), cao hơn so với phương pháp phân lớp dựa vào màu da có giá trị TPR là (0.49 và 0.61). Kết quả này cho thấy, với độ chính xác 76.3255% như trong bài báo này vẫn có thể chấp nhận được.



Hình 6: ví dụ về chức khóa các ảnh khiêu dâm của công cụ

5 KẾT LUẬN VÀ ĐỀ XUẤT

5.1 Kết luận

Trong bày viết này chúng tôi trình bày một hướng tiếp cận trong việc xây dựng công cụ lọc web chạy thời gian thực giúp ngăn chặn sự truy cập đến các trang web chứa thông tin và hình ảnh khiêu dâm, đồi trụy. Từ thực nghiệm cho thấy việc áp dụng kỹ thuật phân lớp văn bản tiếng Việt với SVM cho kết quả phân loại với độ chính xác cao (hơn 90%). Tuy nhiên kỹ thuật này thực sự bị hạn chế về tốc độ tính toán do số chiều vector quá lớn.

Hiện tại công cụ chỉ giới hạn trong việc ngăn chặn các trang web khiêu dâm, tuy nhiên việc mở rộng phạm vi lọc của công cụ sang các chủ đề khác như lọc các trang web có nội dung phản động, bạo lực được thực hiện tương đối dễ dàng (kể cả đối với người dùng không có nhiều kiến thức về khai mỏ dữ liệu) nhờ vào quy trình huấn luyện mô hình tự động.

5.2 Đề xuất

– Cần hoàn thiện proxy, bổ sung các giao thức còn thiếu.

– Cần nghiên cứu áp dụng một giải thuật lựa chọn đặc trưng văn bản thật hiệu quả để thu gọn tập đặc trưng, tăng tốc độ tính toán.

– Nghiên cứu thuật toán giúp kiểm soát download phim và những gói nén.

– Xây dựng thêm bộ lọc văn bản tiếng Anh.

– Nghiên cứu kết hợp trích chọn đặc trưng sift với đặc trưng màu da, huấn luyện mô hình với các tham số của SVM như nghiên cứu của Lopes et al.2009 [18] hoặc Do. 2011 [25] để tăng hiệu quả cho bộ lọc hình ảnh.

TÀI LIỆU THAM KHẢO

1. Christopher D.Manning, Prabhakar Raghavan, Hinrich Schütze , 2009. An Introduction to Information Retrieval. Cambridge University Press Cambridge, England: 117 – 119.
2. David G.Lowe, 2004. "*Distinctive Image Features from Scale-Invariant Keypoints*". International Journal of Computer Vision: 91-110.
3. Jun-Yang, Yu-GangJiang, 1997. "Evaluating Bag-of-Visual-Words Representations in Scene Classification". WOODSTOCK '97 El Paso, Texas, USA: 2 - 3.
4. Thomas Deselaers, Lexi Pimenidis, Hermann Ney, . "Bag-of-Visual-Words Models for Adult Image Classification and Filtering", 2008. 19th International Conference on Pattern Recognition ICPR : 1 – 4.
5. Rongbo Du, Reihaneh Safavi-Naini and Willy Susilo, 2003. "Web Filtering Using Text Classification". The 11th IEEE International Conference on Networks: 325 – 330.
6. W.Hu, H.Zuo, Ou Wu, Yunfei Chen, 2011. "Recognition of Adult Images, Videos, and Web Page Bags". ACM Transactionson Multimedia Computing, Communications and Applications,Vol.7S,No.1: 1 – 28.
7. Mohamed Hammami, Youssef Chahir, Liming Chen, 2003. Combining Text and Image Analysis in the Web filtering System "WEBGUARD". International Conference WWW/Internet: 611 – 618.

8. Youngsoo Kim, Taekyong Nam, Dongho Won, 2006. Text Classification for Harmful Web Document. Computational Science and Its Applications - ICCSA 2006: 545 – 551.
9. Yiming Yang, Jan O. Pedersen, 1997. A comparative Study on Feature Selection in Text Categorization. Proceedings of the 14th international conference on Machine Learning: 412 – 420.
10. Saikat Sen, 2010. Adult Website Classifier. CS229 Machine Learning Course Project, Stanford University, USA. http://cs229.stanford.edu/proj2010/Stanford project CS229.Saikat.Sen.pdf
11. I. Santos, P. Galán-García, A. Santamaría-Ibirika, B. Alonso-Isla, I. Alabau-Sarasola, and Pablo G. Bringas, 2012. Adult Content Filtering through Compression-based Text Classication. CISIS/ICEUTE/SOCO Special Sessions, volume 189 of Advances in Intelligent Systems and Computing: 281-288.
12. Feng Jiao, Wen Gao, Lijuan Duan, Guoqin Cui, 2011. Detecting Adult Image using Multiple Features. Info-tech and Info-net, 2001. Proceedings. ICII 2001 - Beijing. 2001 International Conferences on vol 3: 378 - 383.
13. Zhicheng Zhao, 2010. Combining multiple SVM classifiers for adult image recognition. Network Infrastructure and Digital Content, 2010 2nd IEEE International Conference on: 149 – 153.
14. Đỗ Thanh Nghị, 2012. Khai mỏ dữ liệu. NXB Đại học Cần Thơ.
15. Trần Cao Đệ, Phạm Nguyên Khang, 2012. Phân loại văn bản với Máy học vector hỗ trợ và Cây quyết định. Tạp chí Khoa học 2012 (21a): 52-63.
16. Nguyễn Thị Hoàn, 2010. Phương pháp trích chọn đặc trưng ảnh trong thuật toán Học máy tìm kiếm ảnh áp dụng trong bài toán tìm kiếm sản phẩm. Khóa luận Tốt nghiệp Đại học, Đại học Quốc gia Hà Nội: 13 – 20.
17. Ana P. B.Lopes, Sandra E.F.de Avila, Anderson N. A. Peixoto, Rodrigo S. Oliveira, Marcelo de M. Coelho and Arnaldo de A. Araújo, 2009. Nude Detection in Video using Bag-of-Visual-Features. Computer Graphics and Image Processing (SIBGRAPI): 224 – 231.
18. Ana P. B.Lopes, Sandra E.F.de Avila, Anderson N. A. Peixoto, Rodrigo S. Oliveira, Marcelo de M. Coelho and Arnaldo de A. Araújo, 2009. A Bag-of-Features Approach Based on HUE-SIFT Descriptor for Nude Detection. In Proceedings of the 17th European Signal Processing Conference, Glasgow, Scotland, 2009.
19. Steel, C.M.S, 2012. The Mask-SIFT Cascading Classifier for Pornography Detection. Internet Security (WorldCIS): 139 – 142.
20. Bag of visual words model: recognizing object categories. http://www.robots.ox.ac.uk/~az/icvss08_az_bow.pdf
21. Rob Fergus, 2012. Recognition - Bag of words models http://cs.nyu.edu/~fergus/teaching/vision_2012/9_BoW.pdf
22. L. H. Phương, N. T. M Huyền và V. L. Xuân, 2010. VnTokenizer 4.1.1 – Tách từ tiếng Việt http://vlsp.vietlp.org:8080/demo/?page=resources
23. Chang, C.C., Lin, C.J, 2001. LIBSVM – a library for support vector machines http://www.csie.ntu.edu.tw/~cjlin/libsvm
24. David Lowe, 2005. SiftDemoV4 – SIFT Keypoint Detector http://www.cs.ubc.ca/~lowe/keypoints/
25. T-N.Do, 2011. Detection of Pornographic Images Using Bag-of-Visual-Words and Arcx4 of Random Multinomial Naïve Bayes. 4th International Conference on Theories and Applications of Computer Science, vol.49 of Journal of Science and Technology, Special Issue on Theories and Application of Computer Science: 13 – 24.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

XÂY DỰNG HỆ THỐNG GỢI Ý PHIM DỰA TRÊN MÔ HÌNH NHÂN TỐ LÁNG GIỀNG

Triệu Vĩnh Viêm[1],Triệu Yến Yến[1] và Nguyễn Thái Nghe[2]

[1] *Khoa Công nghệ Thông tin, Trường Đại học Bạc Liêu*
[2] *Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Building a movie recommender system using factor in the neighbors approach

Từ khóa:
Lọc cộng tác, hệ thống gợi ý

Keywords:
Collaborative filtering, recommender system

ABSTRACT

Recommender system can provide suitable items to users by using data about their behavior in the past to predict the future items that users may like. Two successful approaches in recommender system (relying on the collaborative filtering) are the latent factor models which identify potential relationships on both the user and the item; and neighbor models which use similarities between the items or the similarities between the users. In this study, we introduce an approach which is based on the method proposed by Koren (2010) to utilize the advantages of both the aforementioned approaches. Moreover, besides building a web-based movie recommender system, we try to improve the prediction results by adding to the original model several new regularization coefficients for different models' parameters.

TÓM TẮT

Hệ thống gợi ý có thể đưa ra những mục thông tin phù hợp cho người dùng bằng cách dựa vào dữ liệu về hành vi trong quá khứ của họ để đự đoán những mục thông tin mới trong tương lai mà người dùng có thể thích. Hai tiếp cận thành công trong hệ thống gợi ý thuộc vào nhóm lọc công tác là mô hình nhân tố tiềm ẩn - xác định mối quan hệ tiềm ẩn trên cả người dùng và mục thông tin; và mô hình láng giềng - phân tích độ tương tự giữa các mục thông tin với nhau hay giữa những người dùng với nhau. Trong bài viết này, chúng tôi giới thiệu một tiếp cận tích hợp các ưu điểm của cả hai tiếp cận trên dựa vào phương pháp đã được đề xuất bởi Koren (2010). Ở đây, bên cạnh việc xây dưng một hệ thống trên nền web để gợi ý phim ảnh cho người dùng, chúng tôi cũng đã điều chỉnh mô hình đã có bằng cách đưa vào các hệ số regularization trên từng tham số khác nhau của mô hình nhằm cải tiến kết quả dự đoán.

1 GIỚI THIỆU

Sự phát triển của internet đã đưa chúng ta vào thế giới với một lượng lớn các phần tử thông tin như âm nhạc, phim ảnh, sách vở, trang web,… với những đặc tính khác nhau. Kết quả của những thông tin khổng lồ đó, người ta cảm thấy rối rắm và một câu hỏi đặt ra "Cái nào là thích hợp với tôi hơn?" nảy sinh trong tư duy của họ.

May thay, hệ thống gợi ý có thể chỉ ra các thông tin phù hợp trong số thông tin khổng lồ chưa có trật tự, nó sử dụng các kỹ thuật lọc để chọn ra những loại thông tin đặc trưng nhằm hiển thị các phần tử phù hợp với sở thích của người dùng [1]. Theo cách này, hệ thống có tích hợp tính năng gợi ý sẽ thu hút được người dùng cả về sự hài lòng và tin cậy. Các hệ thống gợi ý tiêu biểu như Amazon, Netflix, IMDb, Youtube, Last.fm, MovieLens… đã

tăng được số lượng khách truy cập nhờ vào tính năng hỗ trợ quyết định này của hệ thống.

Hệ thống gợi ý thường dựa trên lọc cộng tác (Collaborative filtering - CF), dựa trên những hành vi quá khứ của người dùng, ví dụ như: lịch sử giao dịch, đánh giá sản phẩm, thời gian xem một mục tin… và đặc biệt là nó không cần thiết phải tạo ra các hồ sơ tường minh (explicit feedback) cho người dùng. Để gợi ý được các mục tin, hệ thống CF cần so sánh các đối tượng cơ bản khác nhau như các mục tin (items) và người dùng (users). Có hai nhánh nghiên cứu chính của lọc cộng tác là tiếp cận láng giềng (neighborhood approach) và các mô hình nhân tố tiềm ẩn (latent factor models).

Hầu hết các tiếp cận chung nhất của CF là dựa trên mô hình lân cận (Neighborhood Models), mô hình user-user (user-based CF) mà được tác giả phân tích rất rõ trong tài liệu [2]. Bên cạnh đó, một tiếp cận trong [3] dựa trên độ tương tự giữa các phần tử (item-based CF) với quy mô tập dữ liệu rất lớn và đưa ra sự gợi ý chất lượng cao trong thời gian thực hiện.

Mô hình nhân tố tiềm ẩn có dạng tương tự như phương pháp phân tích giá trị đơn (Singular Value Decomposition), chuyển đổi cả các mục tin và người dùng vào cùng một không gian tiềm ẩn của các nhân tố, điều này làm chúng có khả năng so sánh trực tiếp. Bên cạnh đó, nhờ vào khả năng biểu diễn và so sánh các khía cạnh dữ liệu khác nhau, tiếp cận này có xu hướng cung cấp kết quả dự đoán cao hơn mô hình láng giềng [4][5]. Tuy nhiên hầu hết các hệ thống thương mại (Amazon, Tivo,…) vẫn còn sử dụng mô hình láng giềng. Sự phổ biến của mô hình này một phần là nhờ vào tính dễ cài đặt và dễ hiểu.

Trong bài báo này, chúng tôi giới thiệu một mô hình dựa trên mô hình đã được Koren [12] đề xuất. Mô hình này có khả năng cải tiến được độ chính xác ngang bằng với mô hình nhân tố tiềm ẩn. Nó không những duy trì được các thuận lợi của mô hình láng giềng, mà còn xử lý được ngay lập tức vấn đề người dùng mới (còn gọi là Cold Start Problem) khi xếp hạng lần đầu mà không cần phải huấn luyện lại. Tiếp cận này được tác giả đặt tên là "Factor in the Neighbors" (Asymmetric SVD – Koren 2008) với khả năng mở rộng có thể tích hợp được nhiều thông tin đầu vào của người dùng (implicit feedback, explicit feedback) và đạt được độ chính xác đáng kể. Bên cạnh việc kế thừa kết quả nghiên cứu của Koren, chúng tôi đã điều chỉnh mô hình đã có bằng cách đưa vào các hệ số chính tắc hóa (regularized) trên từng tham số khác nhau của mô hình nhằm cải tiến kết quả dự đoán. Chúng tôi sẽ ứng dụng mô hình đề xuất này trong việc xây dựng hệ thống gợi ý phim ảnh.

Cấu trúc còn lại của bài báo được trình bày như sau. Chúng tôi bắt đầu với các công việc liên quan trong phần 2, kể cả việc phân tích thuận lợi và bất lợi của hai tiếp cận chính của lọc cộng tác. Sau đó, chúng tôi trình bày một cách điều chỉnh thêm vào cho tiếp cận Asymmetric SVD, mà cho phép tích hợp mượt mà các thông tin đầu vào trong phần 3. Phần 4 sẽ phô bày kết quả thực nghiệm của mô hình Asymmetric SVD chỉ trên quan hệ giữa các mục tin với điều chỉnh regularization trên các tham số khác nhau, dữ liệu sử dụng là MovieLens 100K. Phần 5 sẽ biểu diễn minh họa một website đã được tích hợp gợi ý. Phần 6, chúng tôi sẽ kết luận về kết quả nghiên cứu và nêu hướng phát triển tiếp theo của phương pháp đề xuất.

2 CÁC TIẾP CẬN LIÊN QUAN TRONG HỆ THỐNG GỢI Ý

Hệ thống biểu diễn các đánh giá của người dùng cho các bộ phim qua ma trận **m** người dùng và **n** bộ phim. Chúng tôi dùng các ký tự để phân biệt người dùng và các mục tin: u, v (đại diện cho người dùng), i, j (đại diện cho các bộ phim). Ký hiệu r_{ui} để chỉ mức độ thích của người dùng u cho một bộ phim i nào đó, giá trị này trong khoảng từ 1 đến 5 đối với dữ liệu MovieLens, $\hat{r}_{ui}$ là dự đoán đánh giá của người dùng u cho bộ phim i. Chúng tôi sử dụng dữ liệu MovieLens 100K với 943 người dùng, 1682 bộ phim, mỗi người dùng đánh giá ít nhất 20 bộ phim, như vậy mật độ của ma trận đánh giá chỉ có 6.3%; trường hợp này còn gọi là vấn đề thưa thớt dữ liệu trong lọc cộng tác. Để đối phó với vấn đề "học vẹt" khi dữ liệu thưa, chúng tôi đã sử dụng các hằng số regularization (λ_1, λ_2, λ_3…) trên từng tham số khác nhau. Giá trị tốt nhất của các hằng số này được xác định thông qua nghi thức kiểm tra chéo.

2.1 Ước lượng cơ sở (Baseline Estimates)

Dữ liệu lọc cộng tác tiêu biểu phô bày ảnh hưởng lớn đến người dùng và phần tử, đó là có vài người dùng đánh giá cao hơn những người khác và cho vài phần tử nhận được đánh giá cao hơn những phần tử khác, mà thường được khắc phục bằng ước lượng cơ sở (baseline estimates, biases) sau [6]:

$$b_{ui} = \mu + b_u + b_i \quad (1)$$

Với μ được biết như là trung bình toàn cục các đánh giá của ma trận m xn. Tham số b_u và b_i cho biết độ lệch quan sát được của người dùng u và

phần tử i. Ví dụ, chúng ta muốn ước lượng đánh giá của người dùng "Trung" cho bộ phim "Chúa tể của những chiếc nhẫn" (Lord of the rings - LoR), với trung bình xếp hạng trên tất cả bộ phim là μ=4.5. Mặt khác, trung bình các đánh giá của LoR có xu hướng cao hơn μ là b_i=0.2 và trung bình các đánh giá của "Trung" thấp hơn μ là b_u=0.5. Như vậy, b_{ui} sẽ là 4.2 (4.5-0.5+0.2) với cặp (Trung, LoR).

2.2 Mô hình láng giềng (Neighborhood models)

Nhiều hệ thống vẫn còn tin cậy để sử dụng tiếp cận này, hình thức trước đây của nó là mô hình user-based CF được tác giả phân tích rất rõ ràng trong tài liệu [2]. Phương pháp này dự đoán các đánh giá bằng cách dựa trên việc ghi nhận lại những người dùng có cùng sở thích. Một cách tiếp cận khác cũng dựa trên độ tương tự nhưng làm việc với các mục tin thay vì người dùng, một đánh giá được ước lượng thông qua các đánh giá đã biết của cùng người dùng trên các mục tin tương tự. Item-based CF thì được sử dụng nhiều hơn nhờ vào khả năng mở rộng và cải tiến được độ chính xác của nó [Bell and Koren 2007b, Takács 2007].

Linden .G [3] đã đề cập các kỹ thuật chung nhất của hệ thống gợi ý, rồi phân tích các đặc điểm quan trọng và xác định độ phức tạp về thời gian của chúng. Như kỹ thuật user-based CF, Linden .G cho rằng mô hình dạng này vấp phải vấn đề chi phí thời gian khi gợi ý, do đối mặt với nguồn dữ liệu lớn (hơn 10 triệu khách hàng, hơn 1 triệu phần tử). Ngoài ra, trong [3] tuy sử dụng item-based CF nhưng không xây dựng ma trận item-item, do nhiều cặp phần tử không có những khách hàng chung, hao tốn về thời gian xử lý, cũng như không gian bộ nhớ lưu trữ. Nhóm tác giả đã xây dựng giải thuật tìm các phần tử gợi ý bằng cách tìm độ tương tự của một phần tử i (đã được khách hàng mua hoặc đánh giá – người dùng cần gợi ý) với tập các phần tử i' trong R có liên quan với nó (lịch sử những khách hàng khác đã mua hoặc đánh giá i mà cũng mua i'∈ R), rồi gợi ý các phần tử phổ biến hoặc tương quan nhất. Tiếp cận này gọi tóm tắt là tìm các phần tử mà khách hàng có xu hướng mua cùng nhau.

Các phương pháp láng giềng trở nên phổ biến bởi vì chúng trực quan và dễ dàng liên hệ để cài đặt. Một số đặc tính hữu ích của nó là: Khả năng giải thích (Explainability) – người dùng mong chờ một hệ thống có thể đưa ra lý do cho các gợi ý của nó, khác hơn là phải đối mặt những gợi ý "hộp đen" ("black box"); xử lý được các đánh giá mới, có khả năng cung cấp lời gợi ý ngay lập tức với Item-based CF.

Tuy nhiên, [Bell and Koren 2007b] đã nhấn mạnh vài điểm đáng chú ý về mô hình láng giềng. Một câu hỏi đặt ra của họ là tính phù hợp của độ đo tương tự khi chỉ cô lập cho 2 mục tin mà không phân tích trên một tập đầy đủ các láng giềng. Để khắc phục khó khăn này, các tác giả đã xây dựng một phương pháp láng giềng mới mà cần phải tính các trọng số thêm vào mô hình như sau:

$$\hat{r}_{ui} = b_{ui} + \sum_{j \in S^k(i,u)} \theta_{ij}^u \left(r_{uj} - b_{uj}\right) \qquad (2)$$

Chi tiết hơn về mô hình mới này có thể xem trong tài liệu [7]. Mô hình này được tác giả cải tiến với trọng số quan hệ giữa các phần tử không phụ thuộc vào người dùng cụ thể, cải tiến về độ phức tạp và khả năng tận dụng các thông tin đầu vào với tiếp cận nhân tố trong mô hình láng giềng.

2.3 Mô hình nhân tố tiềm ẩn (Latent factor models)

Tiếp cận phổ biến gần đây của mô hình này là phân giải giá trị đơn (Singular Value Decomposition - SVD) nhờ vào sự hấp dẫn về độ chính xác đạt được và khả năng mở rộng hệ thống. Ngoài ra, phân rã ma trận (Matrix Factorization - MF) hiện đang phát triển rất mạnh với khả năng dự đoán cho lỗi rất thấp.

Đối với mô hình phân rã ma trận cơ bản thì chỉ có một quan hệ giữa 2 thực thể chính cần dự đoán được xem xét, do vậy thiếu độ chính xác dự đoán trong toàn bộ lĩnh vực thực nghiệm [8]. Từ đó, nghiên cứu trong [9] cố gắng cải tiến tiếp cận phân rã ma trận cơ bản với nhiều hướng mở rộng khác nhau của việc cực đại hóa lề trong phân rã ma trận (Maximum Margin Matrix Factorization - MMMF), nghiên cứu trong [10] cũng cố gắng phối hợp thông tin nội dung trực tiếp như là ràng buộc tuyến tính vào tiếp cận phân rã ma trận. Đặc biệt bài báo [11] đã xây dựng một tiếp cận mới cho mô hình nhân tố tiềm ẩn đó là phân rã ma trận đa quan hệ (Multi-Relational Matrix Factorization - MRMF) cho lĩnh vực thực nghiệm là phim ảnh (movies) và gene function prediction.

Mặt khác, các phương pháp mô hình nhân tố tiềm ẩn thường vấp phải vấn đề học lại (relearning) khi có các đánh giá mới (new ratings) [12]. Mô hình dạng này cho kết quả dự đoán có độ chính xác cao nhưng khó đưa ra được lời giải thích cho điều đó.

Trong phần tiếp theo, chúng tôi sẽ giới thiệu phương pháp "Asymmetric SVD với chính tắc hóa

trên từng tham số" dựa trên phương pháp của Koren [12].

3 ASYMMETRIC SVD VỚI CHÍNH TẮC HÓA TRÊN TỪNG THAM SỐ

Với tiếp cận lọc cộng tác, nếu sử dụng phương pháp dựa vào láng giềng thì tùy vào tính quy mô của hệ thống mà ảnh hưởng đến thời gian phản hồi gợi ý, độ chính xác dự đoán chưa cao. Hơn nữa, trong [13] còn đặt một câu hỏi là "tính phù hợp của độ đo tương tự khi nó chỉ cô lập quan hệ của hai phần tử, mà không phân tích sự tương tác bên trong tập hợp đầy đủ các láng giềng", vấn đề người dùng mới cũng tác động đến phương pháp này. Mặt khác, nếu dùng phương pháp dựa trên nhân tố tiềm ẩn thường mất chi phí về thời gian để tìm kiếm tham số, tuy nó giải quyết được vấn đề thưa thớt của dữ liệu và độ chính xác của dự đoán tương đối cao nhưng phải học lại mô hình khi có các đánh giá mới và khó đưa ra được lời giải thích với kết quả dự đoán. Trong [14], Gábor Takács đã đề nghị một phương pháp kết hợp cả Neighborhood Model và Matrix Factorization mà tạo ra hệ thống gợi ý đơn giản, có khả năng mở rộng và đạt được độ chính xác.

Chúng tôi chọn mô hình Asymmetric-SVD được biết như là "Factor in the Neighbors" [12][13] để làm chức năng gợi ý của hệ thống, một phương pháp không những cho phép tận dụng được những ưu điểm của mô hình láng giềng và mô hình nhân tố tiềm ẩn mà còn đạt hiệu quả tốt về thời gian và không gian [12], công thức dự đoán đánh giá như sau:

$$\hat{r}_{ui} = \mu + b_u + b_i + |R(u)|^{-\frac{1}{2}} \sum_{j \in R(u)} (r_{uj} - b_{uj}) q_i^T x_j \quad (3)$$

Tài liệu [13] đã phân tích được nhiều đặc điểm thuận lợi của tiếp cận này. Ngoài ra, tài liệu [12] còn nói rõ hơn sự kết hợp mượt mà của nhiều nguồn thông tin đầu vào từ người dùng bao gồm: phản hồi explicit và implicit.

Công thức (3) ở đây của chúng tôi chỉ thực nghiệm trên việc phân tích mối quan hệ giữa các mục tin bên trong không gian tiềm ẩn. Đặc biệt hơn, chúng tôi đã chính tắc hóa (regularized) trên từng tham số khác nhau trong hàm chi phí như sau:

$$\min_{q_*, x_*, b_*} E = \sum_{(u,i) \in K} \left(r_{ui} - \mu - b_u - b_i - q_i^T |R(u)|^{-\frac{1}{2}} \sum_{j \in R(u)} (r_{uj} - b_{uj}) x_j \right)^2 + \left(\lambda_1 b_u^2 + \lambda_2 b_i^2 + \lambda_3 \|q_i\|^2 + \lambda_4 \sum_{j \in R(u)} \|x_j\|^2 \right) \quad (4)$$

Mô hình này sử dụng L2 – regularization (cụ thể là chuẩn Frobenius) với các hằng số regularization khác nhau (λ_1, λ_2, λ_3, λ_4) để có thể cải tiến được độ chính xác dự đoán so với khi chỉ dùng duy nhất một hằng số cho tất cả các tham số. Sự chính tắc (regularization) làm việc bằng cách thêm vào "hình phạt" liên quan tới tham số trong hàm chi phí của giả thuyết để trừng phạt các trọng số lớn [15].

Hệ thống dựa trên các phản hồi tường minh cố gắng học được các nhân tố thể hiện quan hệ giữa các bộ phim đánh giá trong quá khứ của người dùng và các bộ phim cần dự đoán, trong đó các đánh giá được tập hợp từ MovieLens. Chúng tôi sử dụng giải thuật ***stochastic gradient descent*** để phát hiện quan hệ giữa các bộ phim đã đánh giá trước đây của người dùng và các bộ phim cần dự đoán (mô hình phân rã trọng số quan hệ giữa các bộ phim).

Dưới đây là giải thuật cài đặt cho mô hình huấn luyện tìm các tham số.

FactorizedNeighborhoodModel(Known ratings: r_{ui}, numberOfFactor: f, learnRate: β, Iterations, λ_1, λ_2, λ_3, λ_4)
{
  $q_i \leftarrow N(\mu, \sigma^2)$; $x_j \leftarrow N(\mu, \sigma^2)$
  for(count=1; count<=Iterations; count++) {
    for(u=1; u<=m; u++) {
      $p_u = |R(u)|^{-\frac{1}{2}} \sum (r_{uj} - b_{uj}) x_j$
      sum $\leftarrow$ 0
      for(all i $\in$ R(u)) {
        $\hat{r}_{ui} = \mu + b_u + b_i + q_i p_u^T$;
        $e_{ui} = r_{ui} - \hat{r}_{ui}$;
        sum = sum + $e_{ui} \cdot q_i$;

$$q_i = q_i + \beta \cdot (e_{ui} \cdot p_u - \lambda_3 \cdot q_i)$$
$$b_u = b_u + \beta \cdot (e_{ui} - \lambda_1 \cdot b_u)$$
$$b_i = b_i + \beta \cdot (e_{ui} - \lambda_2 \cdot b_i)$$

```
    } // end all i ∈ R(u)
    for(all i ∈ R(u)) { // tổng bên trong cho x
```

$$x_i = x_i + \beta \cdot \left(|R(u)|^{-\frac{1}{2}} \cdot (r_{ui} - b_{ui}) \cdot \text{sum} - \lambda_4 \cdot x_i \right)$$

```
    }
  } // end all u
} // end steps Iterations
return (qi, xi,bi| i=1,…,n;bu | u=1,…,m
} // end Algorithm
```

Các tham số của mô hình lúc này là q_i, x_i, b_u, b_i, Koren đề xuất phân biệt cách tính các thành phần bias khác nhau [12], khi xét quan hệ giữa các mục tin thì vẫn chọn các bias là hằng số ban đầu, còn các b_u, b_i bên ngoài quan hệ sẽ được tính toán và cập nhật trong quá trình huấn luyện. Với tiếp cận này thì độ phức tạp về thời gian tính toán dự đoán là $0(f - \sum_u |R(u)|)$ và độ phức tạp về không gian tuyến tính với kích cỡ đầu vào 0(m+nf).

4 KẾT QUẢ THỰC NGHIỆM

4.1 Tập dữ liệu huấn luyện

Nghiên cứu này, chúng tôi sử dụng tập dữ liệu MovieLens 100K[1], mỗi người dùng đánh giá ít nhất 20 bộ phim với các tập dữ liệu dùng cho kiểm tra chéo 5-fold sẵn có được chia ra tập train (uX.base) và tập test (uX.test) với X từ 1 đến 5. Chúng tôi tiền xử lý các tập dữ liệu này bao gồm xóa bỏ trường timestamp, trộn ngẫu nhiên các mẫu, chèn vào số mẫu ở đầu mỗi tệp.

4.2 Độ đo

Khi huấn luyện, chúng tôi thử tìm kiếm trên nhiều trường hợp của các siêu tham số (meta-parameters) khác nhau, sử dụng kỹ thuật tìm kiếm lưới (grid search = raw search + smooth search) để đạt được các siêu tham số cho lỗi RMSE trên tập kiểm tra tốt nhất. Lỗi RMSE được xác định bằng công thức:

$$\text{RMSE} = \sqrt{\frac{1}{|D^{\text{test}}|} \sum_{u,i,r \in D^{\text{test}}} (r_{ui} - \hat{r}_{ui})^2} \qquad (5)$$

4.3 Kỹ thuật huấn luyện và kết quả

Chúng tôi sử dụng bước lặp giới hạn (Max_Num_Iters) thay vì lặp cho đến khi hội tụ để quá trình huấn luyện nhanh hơn và giải quyết vấn đề phần tử mới với dự đoán trung bình toàn cục.

Sau đây, chúng tôi thống kê một số kết quả theo từng tốc độ học (LR), số nhân tố (NF) và các regularization (Rb_u, Rb_i, Rq_i, Rx_i). Kết quả thực nghiệm tìm kiếm thô được thể hiện trên 3 tốc độ học: 0.001, 0.005 và 0.01 với số nhân tố 16, 32, 64, cùng với số vòng lặp giới hạn như các bảng sau (trích từ dữ liệu với lỗi RMSE thấp):

Bảng 1: Iters=50, NF=64

LR	Rb_u	Rb_i	Rq_i	Rx_i	RMSE	Time(m)
0.01	0.005	0.005	0.05	0.5	0.9289	0.2290
0.005	0.05	0.05	0.05	0.5	**0.9224**	0.2233
0.001	0.05	0.005	0.005	0.005	0.9318	0.2249

Bảng 2: Iters=100, NF=64

LR	Rbu	Rbi	Rqi	Rxi	RMSE	Time(m)
0.01	0.05	0.005	0.5	0.05	0.9346	0.4467
0.005	0.05	0.005	0.05	0.5	0.9288	0.4456
0.001	0.05	0.005	0.05	0.05	0.9245	0.4449

Bảng 3: Iters=200, NF=64

LR	Rb_u	Rb_i	Rq_i	Rx_i	RMSE	Time(m)
0.01	0.005	0.005	0.5	0.05	0.9347	0.896
0.005	0.005	0.005	0.5	0.05	0.9328	0.892
0.001	0.05	0.05	0.05	0.5	**0.9225**	0.891

Với tiếp cận nhân tố trong mô hình láng giềng [12] số nhân tố cao sẽ cho lỗi dự đoán tốt hơn số nhân tố thấp trên cùng bộ siêu tham số, nên các bảng 1-2-3 thể hiện các giá trị tốt nhất với NF là 64. Sau khi tìm kiếm thô, chúng tôi tiếp tục tìm kiếm mịn hơn trên các giá trị siêu tham số tốt nhất này (dòng tô đen trong Bảng 1), nhằm đạt được độ đo lỗi kiểm tra tốt hơn. Sau đây là kết quả mịn hóa tốt nhất (Bảng 4):

Bảng 4: Tìm mịn

Rb_u	Rb_i	Rq_i	Rx_i	RMSE	Time(m)
0.05	0.0525	0.0475	0.5	**0.92190**	0.2220
0.0475	0.0475	0.0525	0.485	0.92196	0.2179
0.0525	0.0525	0.0475	0.5	0.92199	0.2204
0.0525	0.05	0.0475	0.515	0.92212	0.2202
0.0475	0.0475	0.05	0.515	0.92224	0.2238
0.05	0.0525	0.05	0.485	0.92225	0.2206
0.05	0.05	0.0475	0.485	0.92229	0.2197

Do tính ngẫu nhiên của dữ liệu nên chúng tôi sẽ lặp lại 20 lần việc tính toán kết quả của 8 trường hợp trên rồi lấy kết quả trung bình để tìm được chính xác trường hợp nào là tốt nhất thật sự (Bảng 5).

Bảng 5: Tìm mịn lặp

Rb_u	Rb_i	Rq_i	Rx_i	RMSE	Time(m)
0.0525	0.05	0.0475	0.515	**0.922784**	0.216735
0.05	0.0525	0.05	0.485	0.922829	0.217155
0.05	0.05	0.0475	0.485	0.922841	0.215957
0.0475	0.0475	0.05	0.515	0.922914	0.217372
0.05	0.0525	0.0525	0.5	0.922917	0.216496
0.05	0.0525	0.0475	0.5	0.922925	0.216537
0.0525	0.0525	0.0475	0.5	0.922952	0.218518
0.0475	0.0475	0.0525	0.485	0.922967	0.220743

Thống kê (Bảng 4 và 5) cho ta thấy dữ liệu ngẫu nhiên đã ảnh hưởng ít nhiều đến RMSE (0.9219 và 0.922784).

Các bảng trên là kết quả huấn luyện tìm kiếm siêu tham số trên tập dữ liệu (u1.base, u1.test). Qua đó rõ ràng các siêu tham số regularization khác nhau (Rb_u, Rb_i, Rq_i, Rx_i) sẽ đạt kết quả tốt hơn nếu chỉ dùng cùng một regularization cho tất cả các tham số. Tiếp theo chúng tôi lấy trung bình lỗi trên tất cả các tập dữ liệu của MovieLens 5-fold với cùng các siêu tham số trên tập u1, kết quả cuối cùng như sau:

Bảng 1: Kết quả trên toàn bộ các tập dữ liệu với nghi thức 5-fold

Siêu tham số tìm thô				**Dữ liệu**	**RMSE**
LR=0.005, NF=64, Num_Iter=50				u1	0.9225545
				u2	0.9156835
				u3	0.9087205
Rb_u	Rb_i	Rq_i	Rx_i	u4	0.9106193
0.05	0.05	0.05	0.5	u5	0.9135651
Trung bình					**0.914229**
Siêu tham số tìm mịn				**Dữ liệu**	**RMSE**
LR=0.005, NF=64, Num_Iter=50				u1	0.9228189
				u2	0.9154739
				u3	0.9081727
Rb_u	Rb_i	Rq_i	Rx_i	u4	0.9104163
0.0525	0.05	0.0475	0.515	u5	0.9136031
Trung bình					**0.914097**

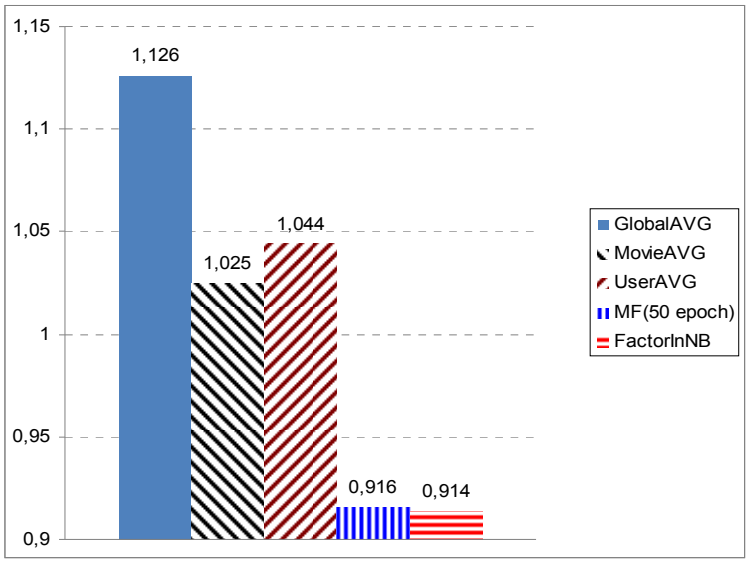
Mỗi tập dữ liệu chúng tôi lặp 5 lần huấn luyện để tính RMSE chính xác

Số liệu trong Bảng 6 đã nói lên được kết quả tìm kiếm mịn có thể cho độ đo RMSE tốt hơn tìm kiếm thô.

4.4 Biểu đồ so sánh

Phần này trình bày so sánh tiếp cận của bài báo với các tiếp cận khác thông qua biểu đồ. Chúng tôi so sánh tiếp cận này với dự đoán trung bình toàn cục (GlobalAVG), trung bình trên bộ phim (MovieAVG), trung bình trên người dùng (UserAVG) và phân rã ma trận có xử lý "cold start problem" (Matrix Factorization - MF) trên cùng tập dữ liệu MovieLens với nghi thức kiểm tra chéo 5-fold.

Dưới đây là biểu đồ biểu diễn độ đo RMSE của các phương pháp đã đề cập.



Hình 1: So sánh các tiếp cận gợi ý

Biểu đồ cho ta thấy lỗi RMSE của tiếp cận Asymmetric SVD với chính tắc hóa trên từng tham số cũng khá tốt hơn phân rã ma trận và tốt hơn nhiều khi so sánh với dự đoán bằng trung bình toàn cục cũng như trung bình trên bộ phim và người dùng.

Thực nghiệm ở đây, chúng tôi chỉ mới xét kết hợp mối quan hệ tiềm ẩn giữa các mục tin, nếu hệ thống được bổ sung thêm nhiều thông tin đầu vào như các phản hồi không tường minh (implicit feedbacks) thì theo [12] RMSE đạt được có thể còn tốt hơn.

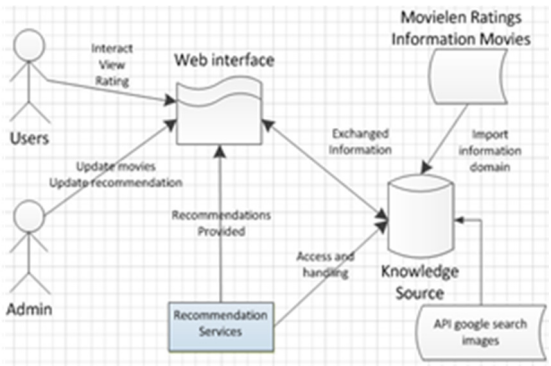
5 XÂY DỰNG HỆ THỐNG

Chúng tôi đã xây dựng một hệ thống gợi ý trên nền web với công nghệ servlets/jsp và cơ sở dữ liệu MySQL.

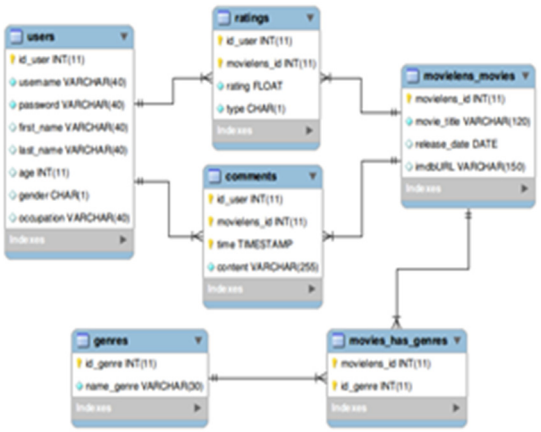
5.1 Kiến trúc tổng quát

Hệ thống rút trích dữ liệu từ tập Movielens 100K và cũng do trong tập dữ liệu này chưa có nhiều nội dung về các bộ phim (đặc biệt là hình ảnh), cho nên nhóm đã sử dụng dịch vụ tìm kiếm ảnh API Google Search Images để tìm dữ liệu cho hệ thống. Người dùng hệ thống có thể chia ra làm 3 nhóm khác nhau: khách vãng lai, người dùng đã đăng ký, người quản trị hệ thống.

Khách vãng lai chỉ được xem và tìm kiếm nội dung thông tin của phim ảnh. Người dùng đã đăng ký sau khi đăng nhập thành công sẽ được tất cả các quyền của khách vãng lai, quyền xếp hạng phim ảnh, bàn luận về chúng và điều đáng nói là họ nhận được các lời gợi ý riêng của mình. Nhóm quản trị hệ thống có thể thêm phim mới và cập nhật được tham số hỗ trợ tính toán gợi ý.



Hình 2: Kiến trúc hệ thống



Hình 3: Mô hình dữ liệu

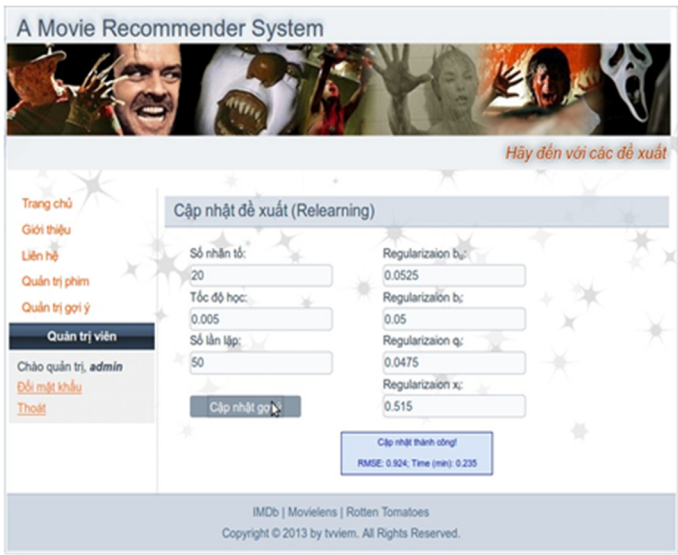
5.2 Mô hình cơ sở dữ liệu

Hệ thống sử dụng MySQL Server Community như là kho dữ liệu. Chúng tôi đã lưu tất cả các tham số sau khi huấn luyện vào cơ sở dữ liệu, những tham số này được dùng để tính toán lời gợi ý ngay sau khi người dùng đăng nhập. Phần tiếp theo sẽ minh họa một số giao diện của hệ thống mà nhóm đã xây dựng.

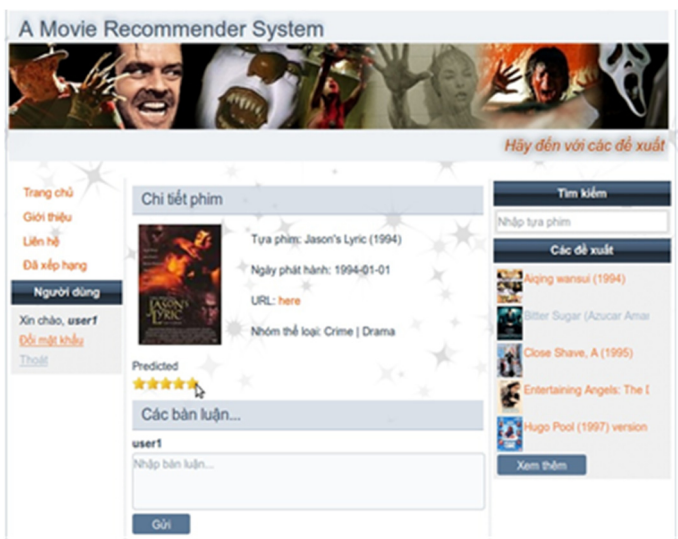
5.3 Các giao diện chính của hệ thống

Người dùng đăng nhập thành công có thể thấy được lời gợi ý cá nhân (nhận thêm đề xuất khác), họ có thể xem chi tiết các bộ phim và xếp hạng lại chúng.

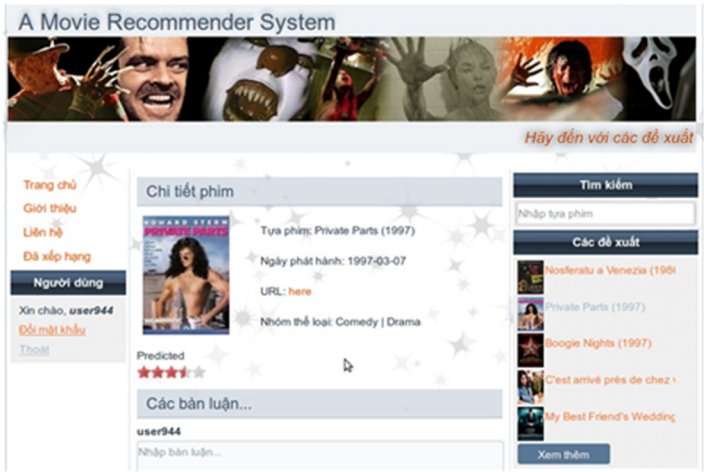
Hơn nữa, hệ thống còn có khả năng giải quyết vấn đề cold start problem với giá trị trung bình toàn cục.



Hình 4:Trang quản trị gợi ý cho phép cập nhật các tham số cho hệ thống (admin)



Hình 5: User đăng nhập thành công, nhận gợi ý, xem chi tiết phim, bàn luận



Hình 6: User944 là người dùng mới được dự đoán trung bình toàn cục và ngẫu nhiên

6 KẾT LUẬN

Phương pháp "Factor in the neighbors" thay vì cung cấp tham số rõ ràng cho những người dùng cụ thể thì nó biểu diễn người dùng thông qua các phần tử mà họ đã đánh giá. Qua kết quả thực nghiệm, ngoài các đặc điểm thuận lợi của tiếp cận này được đề cập trong [12][13], chúng tôi đã xác định được các regularization khác nhau trên từng tham số sẽ cho kết quả tốt hơn thay vì chỉ dùng một regularization cho tất cả các tham số. Bên cạnh đó, chúng tôi đã áp dụng kỹ thuật tìm kiếm lưới cho các siêu tham số để đạt kết quả dự đoán tốt hơn khi chỉ tìm kiếm sơ bộ một vài trường hợp cụ thể nào đó. Nhóm đã xây dựng hệ thống gợi ý nền web với dữ liệu được tập hợp từ Movielens 100K, hệ thống đã thành công trong việc tính toán cũng như cập nhật lời gợi ý trực tuyến. Tuy nhiên, về vấn đề thời gian đăng nhập vẫn còn hạn chế do quá trình nhận các đường dẫn hình ảnh từ dịch vụ API Google Search Images. Vì vậy, tương lai nhóm sẽ lưu trữ dữ liệu hình ảnh trên cơ sở dữ liệu riêng không phụ thuộc vào API google và tiếp tục tận dụng nhiều loại dữ liệu đầu vào của người dùng để dự đoán gợi ý chính xác hơn.

TÀI LIỆU THAM KHẢO

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17(6), (2005).
2. Herlocker, J. L., Konstan, J. A., Borchers, A., Andriedl, J. 1999. An algorithmic framework for performing collaborative filtering. InProceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99). ACM, New York, NY, 230–237.
3. G. Linden, B. Smith, and J. York, 2003. Amazon.com recommendations: Item - item collaborative filtering. IEEE Internet Comput. 7, 1, 76–80.
4. Gábor Takács, István Pilászy, Bottyán Németh: Major components of the Gravity Recommendation System, Volume 9, Issue 2, p80-83, 2007.
5. Arkadiusz Paterek: Improving regularized singular value decomposition for collaborative filtering, KDDCup.07 August 12, 2007, San Jose, California, USA.
6. Yehuda Koren, Yahoo Research, Robert Bell and Chris Volinsky, AT&T Labs - Research: "Matrix Factorization techniques for recommender systems", Published by the IEEE Computer Society, 2009.
7. Bell, R.M. and Koren, Y. 2007b. Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In Proceedings of the IEEE International Conference on Data Mining (ICDM). IEEE Computer Society, 43–52.
8. Gábor Takács, István Pilászy, Bottyán Németh: "Investigation of Various Matrix Factorization Methods for Large Recommender Systems", 2nd Netflix-KDD Workshop, August 24, 2008, Las Vegas, NV, USA.
9. Markus Weimer, Alexandros Karatzoglou, Alex Smola: "Improving maximum margin matrix factorization", Mach Learn (2008) 72: 263–276.

10. Peter Forbes, Mu Zhu: “Content-boosted Matrix Factorization for Recommender Systems: Experiments with Recipe Recommendation”, RecSys’11, October 23 - 27, 2011, Chicago, Illinois, USA.
11. Christoph Lippert, Stefan Hagen Weber, Yi Huang, Volker Tresp, Matthias Schubert, Hans-Peter Kriegel: “Relation Prediction in Multi-Relational Domains using Matrix-Factorization”, NIPS 2008 workshop on "Structured Input, Structured Output" (SISO 2008).
12. Koren, Y. (2010)”: “Factor in the neighbors: Scalable and accurate collaborative filtering.”, AT&T Labs - Research 180 Park Ave, Florham Park, NJ 07932.
13. Y. Koren: “Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model”, Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008.
14. Gábor Takács, István Pilászy, Bottyán Németh: “Investigation of Various Matrix Factorization Methods for Large Recommender Systems”, 2nd Netflix-KDD Workshop, August 24, 2008, Las Vegas, NV, USA.
15. Aleks Jakulin: “Machine Learning Based on Attribute Interactions”, University of Ljubljana, Se˘zana, June 13, 2005.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

ĐIỀU KHIỂN TẮC NGHẼN TRONG GIAO THỨC TRUYỀN ĐA ĐƯỜNG CHO CÁC ỨNG DỤNG MULTIMEDIA

Lê Phong Dũ[1] và Lê Tuấn Anh[2]

[1] *Ban Phát triển Hệ thống Công nghệ Thông tin, Trường Đại học Trà Vinh*
[2] *Khoa Công nghệ Thông tin, Trường Đại học Thủ Dầu Một*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Congestion control in multipath protocol for multimedia applications

Từ khóa:
Thuật toán điều khiển tắc nghẽn, TFRC, Multipath TCP

Keywords:
Congestion control algorithm, TFRC, Multipath TCP

ABSTRACT

Recently, multipath transport control protocol (multipath TCP) is introduced as an extended protocol with added features from single-path TCP. It allows data packets of a multipath TCP connection spreadded on several paths simultaneously. A congestion control algorithm for multipath TCP should meet three goals: (i) throughput improvement; (ii) fair sharing to existing TCP flows; (iii) congestion balance. However, data rate of MPTCP has high variation that not suitable for multimedia applications which requires smooth data rate.

In this paper, we propose an extension algorithm of single path TFRC, named MPTFRC, designed from both the analytical model of TCP Reno at flow level and the technique of fluctuation prevention between paths at packet level. The simulation results demonstrate that the proposed MPTFRC algorithm not only satisfies the three goals but also provides data rate smoother than that of MPTCP while preserving fair sharing to the existing TCP Reno and MPTCP flows.

TÓM TẮT

Gần đây, giao thức truyền đa đường TCP (multipath TCP) được đề xuất như là một giao thức mở rộng với một số đặc điểm thêm vào từ giao thức đơn đường TCP. Nó cho phép các gói dữ liệu một kết nối multipath TCP được truyền tải trên nhiều đường (path) đồng thời. Một thuật toán điều khiển tắc nghẽn cho giao thức truyền đa đường TCP phải thỏa mãn ba tiêu chí: (i) nâng cao thông lượng truyền; (ii) đảm bảo tính công bằng đối với các luồng TCP hiện có; (iii) cân bằng tắc nghẽn. Tuy nhiên, tốc độ truyền dữ liệu của MPTCP biến thiên lớn, không phù hợp cho các ứng dụng multimedia mà chúng yêu cầu tốc độ dữ liệu mượt.

Trong báo cáo này, chúng tôi đề xuất một thuật toán mở rộng của TFRC đơn đường, được đặt tên là MPTFRC, được thiết kế xuất phát từ mô hình phân tích của TCP Reno gốc tại mức luồng và kỹ thuật tránh đánh võng giữa các đường ở mức gói. Các kết quả mô phỏng đã chứng tỏ rằng thuật toán đề xuất MPTFRC không những đảm bảo ba tiêu chí trên mà còn cung cấp tốc độ dữ liệu mượt hơn MPTCP trong khi vẫn duy trì được sự chia sẻ công bằng với các luồng TCP Reno và MPTCP hiện có.

1 GIỚI THIỆU

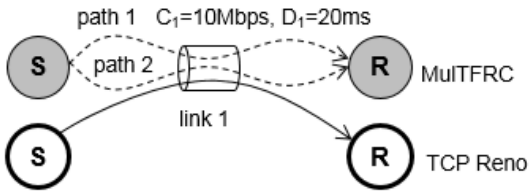
Kiến trúc multipath TCP [8] là giao thức mở rộng các đặc điểm từ giao thức TCP, cho phép một kết nối phân chia thành nhiều đường (path) và dữ liệu được truyền trên các đường đồng thời. Multipath TCP hoạt động giống như TCP và mở rộng thêm các API nhằm cung cấp thêm chức năng điều khiển cho các ứng dụng multipath TCP.

Một kết nối multipath TCP là một tập hợp của nhiều subflow mà mỗi subflow điều khiển một đường, sử dụng cửa sổ điều khiển tắc nghẽn để điều chỉnh tốc độ trên mỗi đường. Gần đây, thuật toán điều khiển tắc nghẽn MPTCP [5] đã được IETF phê duyệt hoạt động trên kiến trúc multipath TCP [8]. MPTCP thực hiện điều khiển tốc độ gửi dữ liệu phối hợp giữa các đường đảm bảo ba tiêu chí: *(i) nâng cao thông lượng truyền, (ii) đảm bảo tính công bằng đối với các luồng TCP hiện có và (iii) có khả năng cân bằng tắc nghẽn*: chuyển dữ liệu từ đường có tắc nghẽn nhiều sang đường có tắc nghẽn ít hơn. Vì MPTCP được thiết kế cho đa ứng dụng, nên tốc độ truyền dữ liệu của MPTCP biến thiên lớn, không phù hợp cho các ứng dụng multimedia mà ở đó chúng yêu cầu *tốc độ dữ liệu truyền mượt (chúng tôi gọi yêu cầu này tiêu chí (iv)* cho các ứng dụng multimedia).

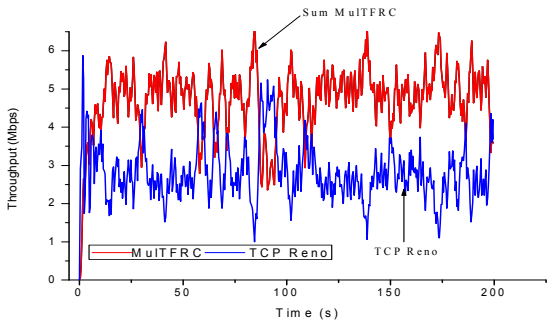
Đã có nhiều nghiên cứu điều khiển tắc nghẽn cho các ứng dụng multimedia trong giao thức đa đường, cụ thể: MultiTCP [12] là một điều khiển định hướng bên nhận (receiver-driven control), nó tăng tốc độ như TCP đơn đường và giảm tốc độ tỉ lệ nghịch với bình phương của số lượng đường được sử dụng. Ngược lại, DMP scheme [10] là một điều khiển định hướng bên gửi (sender-driven control), nó không quan tâm các luồng con có chia sẻ đường truyền cổ chai hay không, cả hai giải pháp trên không quan tâm đến tốc độ mượt và cân bằng tắc nghẽn. Ngoài ra, TCP-ROME [11] giảm sự biến động của tốc độ tức thời bằng cách sử dụng điều khiển dựa trên tốc độ như trong TFRC [3], đầu nhận trong TCP-ROME điều chỉnh tốc độ của luồng con tùy theo tốc độ video yêu cầu và tỉ lệ thông lượng của mỗi luồng, không quan tâm có đường truyền cổ chai.

Gần đây, MulTFRC [1] đã được đề xuất. Nó là phiên bản mở rộng của TFRC [3] vốn là giao thức đơn đường và được thiết kế cho các ứng dụng multimedia. Vì MulTFRC được thiết kế với giả định rằng các đường có cùng round-trip-time (RTT). Trên thực tế, giả định này hiếm xảy ra, cho nên MulTFRC không đảm bảo thỏa mãn tiêu chí trong giao thức đa đường, do nó không có khả năng bù khác biệt RTT, mặc dù nó có khả năng cung cấp tốc độ truyền dữ liệu mượt.

Để minh họa sự chia sẻ không công bằng giữa hai luồng của MulTFRC với một luồng đơn TCP Reno tại một đường truyền nút cổ chai. Mô hình mạng mô phỏng như Hình 1.



Hình 1: Mô hình mạng đánh giá chia sẻ công bằng của MulTFRC



Hình 2: Thông lượng chia sẻ công bằng giữa MulTFRC và TCP Reno

Hình 2 cho thấy rằng, tổng thông lượng của MulTFRC gấp đôi so với thông lượng của TCP Reno. Khi chia sẻ công bằng với các luồng TCP đơn đường. Một kết nối trong giao thức đa đường chia sẻ công bằng với kết nối trong giao thức đơn đường khi nhiều luồng con của giao thức đa đường cạnh tranh với nhau. Do đó, thuật toán MulTFRC không thỏa mãn tiêu chí thứ hai về chia sẻ công bằng tại đường truyền cổ chai.

Trong bài báo này, chúng tôi đề xuất một thuật toán mở rộng của TFRC [3] đơn đường, được đặt tên là MPTFRC, xuất phát từ mô hình phân tích của TFRC gốc [3] và kết hợp với sự cải tiến để tránh đánh võng giữa các đường ở mức gói. Các kết quả mô phỏng đã chứng tỏ rằng thuật toán đề xuất MPTFRC không những đảm bảo ba tiêu chí trên mà còn cung cấp tốc độ dữ liệu mượt hơn MPTCP trong khi vẫn duy trì được sự chia sẻ công bằng với TCP chuẩn, TFRC và MPTCP hiện có.

Phần còn lại của bài báo chúng tôi tổ chức như sau: Chúng tôi mô tả chi tiết thiết kế thuật toán MPTFRC trong Phần II. Chúng tôi đánh giá hiệu

năng của thuật toán trong Phần III. Và kết luận sẽ trình bày trong Phần IV.

2 THIẾT KẾ GIAO THỨC MPTFRC

Trong phần này chúng tôi trình bày giao thức đề xuất MPTFRC mở rộng từ giao thức TFRC [3]. Trước tiên, chúng tôi tiến hành mở rộng thuật toán điều khiển tắc nghẽn TFRC đơn đường thành đa đường có sự phối hợp giữa các đường. Sau đó, kết hợp với điều chỉnh tránh đánh võng để đảm bảo thuật toán hoạt động hiệu quả trong các điều kiện mạng khác nhau. Hiện tượng đánh võng sẽ không thể nhìn thấy trong lúc thiết kế đa đường (nghĩa là quá trình thiết kế chỉ thực hiện ở mức luồng) mà chỉ có thể thấy được ở mức gói. Giả sử có hai đường có cùng điều kiện mạng (cùng RTT, băng thông và cùng mức độ nghẽn mạng), sự đánh võng được mô tả là: dữ liệu gần như chỉ truyền trên một đường trong khoảng thời gian ngẫu nhiên và sau đó chỉ truyền trên đường còn lại và lặp lại như thế [2].

Để cải tiến công thức truyền tốc độ từ thuật toán TFRC gốc [3], việc chứng minh công thức dựa vào giao thức TCP đa đường. Xuất phát từ công thức điều khiển tốc độ trong TFRC gốc được đề xuất trong [3].

$$B(p) \approx \frac{1}{RTT\sqrt{\frac{2pb}{3}} + T_0 \min\left(1, 3\sqrt{\frac{3pb}{8}}\right) p\left(1+32p^2\right)},$$

trong đó, công thức tính thông lượng $B(p)$ được tính bằng gói tin/giây, RTT là round-trip-time, tỉ lệ mất gói tin là p, b là số gói tin được xác nhận bởi một ACK ($b = 1$), T_0 thời gian truyền lại hết thời gian chờ (RTO). Giá trị $T_0 = 4 \times RTT$. Chúng tôi giả sử rằng R là tập hợp tất cả các đường, $r \in R$ là một đường cụ thể. Trọng số đề xuất điều chỉnh phối hợp tốc độ truyền dữ liệu đưa vào giao thức MPTFRC ở trạng thái ổn định là

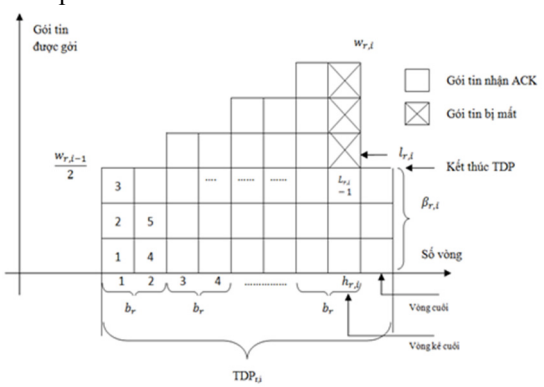
$$\alpha_r = \left(\frac{\hat{x}_r}{\sum \hat{x}_r}\right)^2,$$

Trong đó $\hat{x}_r$ là tốc độ truyền dữ liệu ở trạng thái ổn định, gọi b_r là số gói tin được xác nhận bởi một ACK trên đường r .

Trên mỗi đường r , giả sử w_r là kích thước cửa sổ tránh tắc nghẽn. Cơ chế điều khiển tránh tắc nghẽn của các đường r được thực hiện theo các vòng truyền tin, một vòng bắt đầu bằng việc gửi w_r gói tin trong cửa sổ tắc nghẽn và trong quá trình các gói tin được gửi đi sẽ không có bất kỳ gói tin nào được gửi, trước khi ACK đầu tiên được gửi về. Khi ACK đầu tiên được nhận, đánh dấu vòng đầu tiên kết thúc và chuyển sang vòng thứ 2. Kích thước cửa sổ ở vòng thứ 2 $w_r^{'} = w_r + \frac{a_r}{b_r}$, kích thước cửa sổ sẽ tăng tuyến tính với chu kỳ $\frac{a_r}{b_r}$ trên một vòng hay RTT_r trong giai đoạn tránh tắc nghẽn và không mất gói tin. Khoảng thời gian của một vòng bằng với RTT_r và gói tin bị mất trong một vòng thì độc lập với các vòng khác.

2.1 Phát hiện mất gói tin theo cơ chế dupACK

Xét trên mỗi đường r , chúng tôi định nghĩa thông lượng $B_r(p)$ được xác định số gói tin được truyền đi trong khoảng thời gian nhất định và TDP_r là khoảng giữa hai lần xảy ra mất gói tin trong cơ chế dupACK.



Hình 3: Cơ chế phát hiện mất gói tin dạng dupACK

Trong một TDP_r thứ i , thì $Y_{r,i}$ là số gói tin được gửi đi, $A_{r,i}$ là khoảng thời gian và $w_{r,i}$ là kích thước cửa sổ. Mối quan hệ giữa thông lượng và xác suất mất gói tin được tính như sau:

$$B_r(p) = \frac{E[Y_r]}{E[A_r]}$$

Giả sử hệ số $l_{r,i}$ là gói tin bị mất đầu tiên trong TDP_r thứ i tại vòng $h_{r,i}$. Tổng $Y_{r,i} = l_{r,i} + w_{r,i} - 1$ gói tin sẽ được thêm vào vòng $h_{r,i} + 1$ theo công thức:

$$E\left[Y_r\right] = E\left[L_r\right] + E\left[W_r\right] - 1$$

Do $l_{r,i}$ trong vòng độc lập với các gói tin trong các vòng khác, vì thế $\{L_{r,i}\}_i$ là một chuỗi các biến ngẫu nhiên độc lập và phân phối giống nhau. Xác suất $l_{r,i} = k_r$ tức là $k_r - 1$ gói tin đã gửi thành công trước khi gói tin bị mất xảy ra.

$P\left[l_{r,i} = \mathrm{k_r}\right] = (1-p_r)^{k_r-1} p_r, k_r = 1,2,3,\ldots$ tức là $l_{r,i}$ được xác định là:

$$E\left[L_{r,i}\right] = \sum_{k_r=1}^{\infty} (1-p_r)^{k_r-1} p_r k_r = \frac{1}{p_r}$$

Đặt $t_{r,ij}$ là khoảng thời gian của vòng thứ j trong TDP$_r$ thứ i. Khi đó, thời gian của một TDP$_r$ thứ i là $A_{r,i} = \sum_{j=1}^{h_{r,i}+1} t_{r,ij}$. $t_{r,ij}$ là biến ngẫu nhiên và độc lập với kích thước cửa sổ, vì thế nó độc lập với vòng thứ j.

$$E\left[A_r\right] = \left(E\left[H_r\right] + 1\right) E\left[t_r\right],$$

trong đó biểu diễn $RTT_r = E[t_r]$.

Cuối cùng xác định được các công thức sau:

$$E\left[H_r\right] = \frac{2\alpha_r + b_r}{6\alpha_r} + \sqrt{\left(\frac{2\alpha_r + b_r}{6b_r}\right)^2 + \frac{2b_r(1-p_r)}{3\alpha_r p_r}}$$

$$E\left[A_r\right] = RTT_r\left(\frac{2\alpha_r + b_r}{6\alpha_r} + \sqrt{\left(\frac{2\alpha_r + b_r}{6b_r}\right)^2 + \frac{2b_r(1-p_r)}{3\alpha_r p_r}} + 1\right)$$

$$B_r(p) = \frac{1}{RTT_r}\sqrt{\frac{3\alpha_r}{2b_r p_r}} + O\left(\frac{1}{\sqrt{p_r}}\right)$$

2.2 Phát hiện mất gói tin theo cơ chế dupACK và Time-out

Đặt Z_r^{TO} là khoảng thời gian trong trường hợp mất gói tin dưới dạng time-out và Z_r^{TD} khoảng thời gian trong trường hợp mất gói tin dưới dạng dupACK. S_r là tổng thời gian được định nghĩa như sau:

$$S_r = Z_r^{TO} + Z_r^{TD}.$$

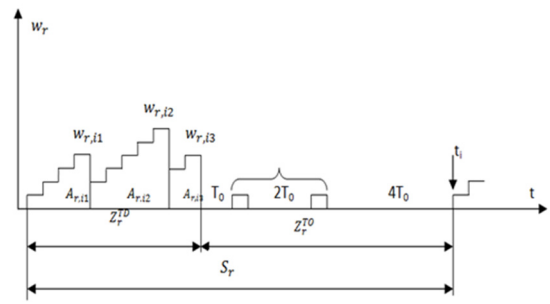
Đặt M_r là số gói tin đã truyền trong khoảng thời gian S_r, khi đó $\{S_r, M_r\}$ là một chuỗi các biến ngẫu nhiên. Thông lượng được tính như sau:

$$B_r(p) = \frac{E[M_r]}{E[S_r]}$$

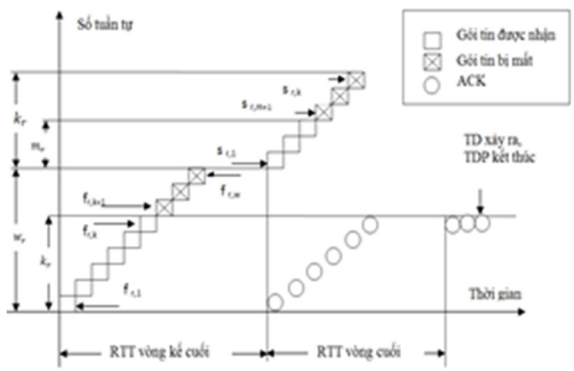
Đặt n_r là số TDP$_r$ trong khoảng thời gian Z_r^{TD}. Trong TDP$_r$ thứ i, xác định các giá trị $Y_{r,i}$ là số gói tin được gửi, $A_{r,i}$ là khoảng thời gian và $W_{r,i}$ là kích thước cửa sổ. Tương tự, R_r là số gói tin gửi trong khoảng thời gian Z_r^{TO}.

Xét trong khoảng thời gian Z_r^{TD} có n_r TDP$_r$ thì có $n_r - 1$ TDP$_r$ kết thúc dưới dạng dupACK và TDP$_r$ cuối cùng kết thúc dưới dạng time-out [9]. Khi đó $\mathbf{Y}_{r,i}$ và $\mathbf{A}_{r,i}$ không phụ thuộc vào time-out, có nghĩa là nó sử dụng lại công thức tính dạng dupACK. Thông lượng cụ thể được tính như sau:

$$B_r(p) = \frac{E\left[Y_r\right] + Q_r \times E\left[R_r\right]}{E\left[A_r\right] + Q_r \times E\left[Z_r^{TO}\right]}$$



Hình 4: Mất gói tin dạng dupACK và time-out



Hình 5: Gói tin bị mất trong vòng kế cuối

Xét các gói tin từ $\mathbf{f}_{r,1} \ldots \mathbf{f}_{r,w}$ được gửi tại vòng kế cuối, các gói $\mathbf{f}_{r,1} \ldots \mathbf{f}_{r,k}$ nhận được ACK và giả

sử gói tin $f_{r,k+1}$ là gói tin đầu tiên bị mất. Do đó, tất cả các gói tin từ $f_{r,k+1}$ trong vòng kế cuối đều bị mất. Tuy nhiên, khi các gói tin $f_{r,1} \dots f_{r,k}$ nhận được ACK thì có k_r gói tin khác $s_{r,1} \dots s_{r,k}$ được gửi trong vòng kế tiếp mà có thể được xem là vòng cuối cùng có một số gói tin bị mất, giả sử gói tin bị mất là $s_{r,m+1}$, tức là gói tin từ $s_{r,m+2} \dots s_{r,k}$ bị mất trong vòng cuối cùng [9].

Đặt m_r là số gói tin đã gửi thành công trong vòng cuối cùng được xác nhận bởi ACK cho các gói tin $f_{r,k}$. Giả sử $A(w_r, k_r)$ là xác suất mà k_r gói tin đầu tiên nhận được ACK trong một vòng của w_r gói tin thì:

$$A\left(w_r, k_r\right) = \frac{\left(1 - p_r\right)^{k_r} + p_r}{1 - \left(1 - p_r\right)^{w_r}}$$

Giả sử $C(n_r, m_r)$ là xác suất mà m_r gói tin nhận được ACK trong vòng cuối cùng (n_r là gói tin được gửi) và phần còn lại của những gói tin trong vòng. Nếu có bị mất gói tin thì

$$C\left(n_r, m_r\right) = \begin{cases} \left(1 - p_r\right)^{m_r} + p, m_r \le n_r - 1 \\ \left(1 - p_r\right)^{m_r}, m_r = n_r \end{cases}$$

Đặt $\widetilde{Q_r}(w_r)$ là xác suất mất gói tin trên cửa sổ với kích thước w_r dưới dạng time-out.

Xác suất có gói tin $f_{r,k+1}$ bị mất trong vòng kế cuối và gói tin bị mất $s_{r,m+1}$ trong vòng cuối cùng và có giá trị $\min\left(1, \frac{3}{w_r}\right)$.

Tiếp theo, xem xét khi một gói tin được truyền giữa hai chuỗi time-out liên tiếp. Một chuỗi k_r dạng time-out xảy thì có $k_r - 1$ liên tiếp mất gói tin. Do đó:

$$P\left[R_r = k_r\right] = p_r^{k_r - 1}\left(1 - p_r\right)$$

R_r được tính như sau:

$$E\left[R_r\right] = \sum_{k_r=1}^{\infty} k_r p_r \left[R_r = k_r\right] = \frac{1}{1 - p_r}$$

Khoảng thời gian trung bình của chuỗi time-out bao gồm cả việc truyền lại. Thấy rằng 6 time-out đầu tiên kết thúc trong 1 chuỗi có chiều dài là $2^{i-1}T_0, i = 1 \dots 6$ với time-out trung bình theo sau có chiều dài $64T_0$ [9]. Khoảng thời gian của một chuỗi với k_r time-out kết thúc được tính như sau:

$$G_{k_r} = \begin{cases} \left(2^{k_r} - 1\right)T_0, k \le 6 \\ \left(63 + 64\left(k_r - 6\right)\right)T_0, k \ge 7 \end{cases}$$

$$E\left[Z_r^{TO}\right] = \sum_{k_r=1}^{\infty} G_{k_r} p_k \left[R_r = k_r\right]$$

$$E\left[Z_r^{TO}\right] = T_0 \frac{1 + p_r + 2p_r^2 + 4p_r^3 + 8p_r^4 + 16p_r^5 + 32p_r^6}{1 - p_r}$$

$$B_r(p) = \frac{1}{RTT_r\sqrt{\frac{2b_r p_r}{3\alpha_r}} + T_0 \min\left(1, 3\sqrt{\frac{3b_r p_r}{8\alpha_r}}\right) p_r\left(1 + 32p_r^2\right)}$$

2.3 Thuật toán MPTFRC

Trong quá trình thiết kế MPTFRC, chúng tôi nhận thấy rằng, tốc độ truyền dữ liệu của MPTFRC nếu chỉ cập nhật tăng, giảm theo công thức (1) thì xảy ra hiện tượng đánh võng. Hiện tượng này được mô tả như sau:

Giả sử rằng MPTFRC truyền dữ liệu trên hai đường độc lập có cùng băng thông, độ trễ và cùng mức độ nghẽn mạng thì sự đánh võng được mô tả là: dữ liệu gần như chỉ truyền trên một đường trong khoảng thời gian ngẫu nhiên và sau đó chỉ truyền trên đường còn lại và lặp lại như thế. Hiện tượng này cũng xuất hiện trong thiết kế MPTCP [8]. Để tránh hiện tượng này, chúng tôi sử dụng kỹ thuật được đề xuất trong [7] là thêm tham số ε_r trên đường r. Thuật toán tăng tốc độ truyền trên đường r được trình bày:

Thuật toán điều khiển tăng tốc độ trên đường r trong giao thức MPTFRC:

```
if (curr_rate_r<target_rate_r)
{
 mult=(now-last_change)/rtt_r
 if (mult > 2) mult=2;

curr_rate_r=curr_rate_r+(size_r/rtt_r
) × mult × α_r + ε_r × size_r/rtt_r ;
 }
```

```
last_change_ = now;
```

Tham số ε_r được thêm vào công thức (1) để tránh đánh võng được thực hiện như sau:

Đặt $\mathbf{RTT}_r$ và w_r là round-trip-time và kích thước cửa sổ trên đường r. Tiếp theo cần xác định đường có kích thước cửa sổ lớn nhất và đường tốt nhất. Công thức xác định được định nghĩa như sau:

$$M = \left\{ i(t) \middle| i(t) = \arg\max_{r \in R} w_r \right\}$$

$$B = \left\{ j(t) \middle| j(t) = \arg\max_{r \in R} \frac{1}{RTT_r^2 (p_r)} \right\}$$

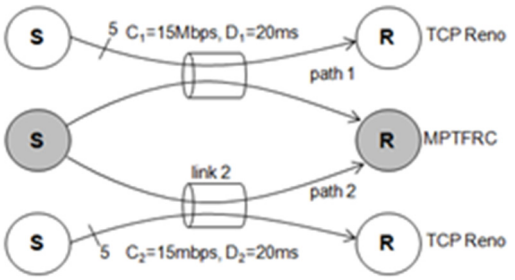
M là tập hợp những đường với kích thước cửa sổ lớn nhất, B là tập hợp những đường mà được xem xét là đường tốt nhất. Để đảm bảo không đánh võng trong quá trình truyền dữ liệu, sử dụng tham số ε_r tránh đánh võng trên từng đường như sau:

$$\varepsilon_r = \begin{cases} \dfrac{\frac{1}{|R|}}{|B \backslash M|} \text{ nếu } r \in B \backslash M \neq \emptyset \\ -\dfrac{\frac{1}{|R|}}{|M|} \text{ nếu } r \in M \text{ và } B \backslash M \neq \emptyset \\ 0 \qquad \text{ngược lại} \end{cases}$$

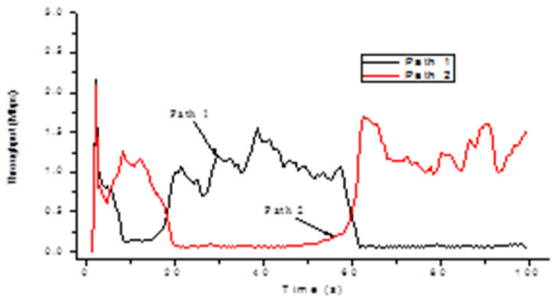
Trong đó, $B \backslash M$ là tập hợp những phần tử có trong B nhưng không có trong M, $\emptyset$ là phần tử rỗng và $|R|$ là tổng số đường truyền dữ liệu. Khi đó $\sum_{r \in R} \varepsilon_r = 0$.

Nếu đường tốt nhất có kích thước cửa sổ lớn nhất, thì $\varepsilon_r = 0$ bất kỳ với $r \in R$. Tuy nhiên, nếu đường tốt nhất có kích thước cửa sổ nhỏ, tức là nếu $B \backslash M \neq \emptyset$ thì ε_r là một số dương nếu $r \in B \backslash M$, ε_r là số âm nếu $r \in M$.

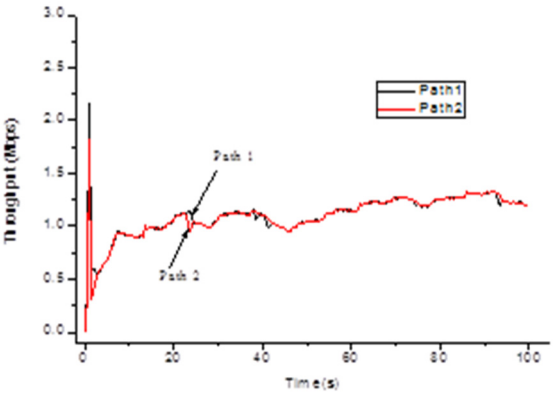
Chúng tôi sẽ mô phỏng sự đánh võng khi không có tham số ε_r và tránh đánh võng bằng tham số ε_r. Mô phỏng được thực hiện như Hình 6. Ở đó có 5 luồng TCP Reno trên các link 1 và 2 để tạo lưu lượng nền (background traffic).



Hình 6: Mô hình mạng đánh giá đánh võng



Hình 7: không sử dụng tham số ε_r



Hình 8: sử dụng tham số ε_r

Trong Hình 7 khi truyền dữ liệu chúng tôi không sử dụng tham số tránh đánh võng ε_r. Thông lượng truyền trên hai đường thay đổi. Đầu tiên thông lượng tăng nhanh trên đường truyền thứ 2 (path 2) sau đó bắt đầu giảm tốc độ, thông lượng truyền thay đổi chỉ truyền trên đường 1 (path 1) và gần như chỉ truyền trên một đường trong khoảng thời gian ngẫu nhiên, tốc độ truyền trên hai đường luân phiên thay đổi, hiện tượng đánh võng xảy ra

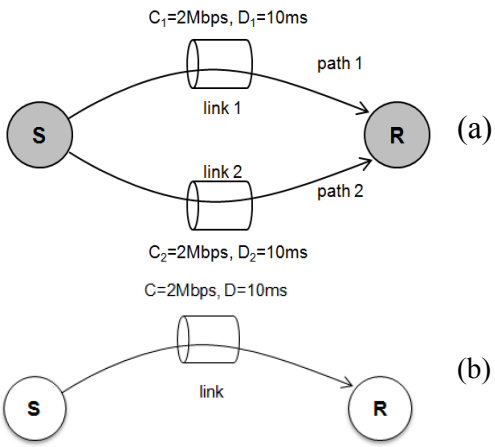
trên hai đường. Khi sử dụng tham số tránh đánh võng ε_r thì không xuất hiện sự đánh võng như trong Hình 8, ở đó tốc độ trên hai đường truyền (path 1 và path 2) gần như bằng nhau trong suốt 100 giây mô phỏng.

3 ĐÁNH GIÁ HIỆU NĂNG

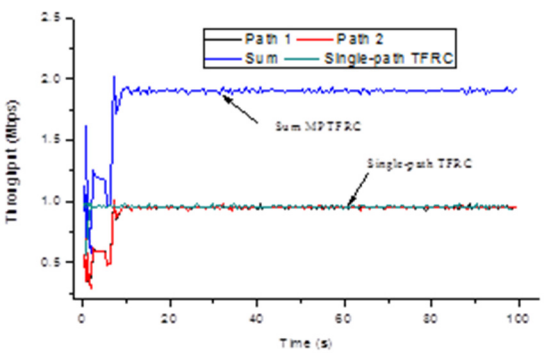
Để đánh giá thuật toán MPTFRC đã đề xuất trong phần 2, chúng tôi tiến hành mô phỏng bằng bộ công cụ mô phỏng mạng NS2-31 [6], chạy trong môi trường hệ điều hành Ubuntu 10.4. Với các tham số mô phỏng như sau: kích thước gói dữ liệu là 576 byte, sử dụng cơ chế hàng đợi active RED, tốc độ lấy mẫu sau mỗi 500 ms, kích thước bộ đệm bằng băng thông nhân với độ trễ.

a. Đánh giá theo tiêu chí tăng thông lượng

Để đánh giá tiêu chí tăng thông lượng của MPTFRC, đầu tiên chúng tôi mô phỏng hai đường MPTFRC qua hai link như Hình 9(a), sau đó chúng tôi tiếp tục mô phỏng riêng biệt một luồng TFRC qua 1 link như Hình 9(b).



Hình 9: Mô hình mạng đánh giá tiêu chí tăng thông lượng

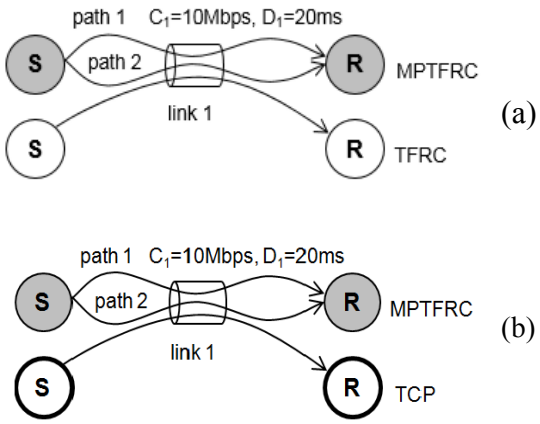


Hình 10: Tăng thông lượng của MPTFRC

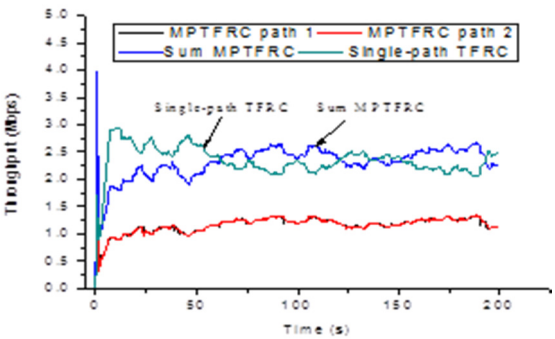
Kết quả Hình 10 thấy rằng, thông lượng tổng hai đường của MPTFRC là gần gấp đôi so với TFRC đơn đường, tăng thông lượng đạt được như trong thuật toán điều khiển tắc nghẽn giao thức truyền đa đường. Việc tăng thông lượng của MPTFRC so với TFRC đơn đường phụ thuộc vào số lượng Subflow trong một kết nối MPTFRC.

b. Đánh giá theo tiêu chí chia sẻ công bằng

Trong phần này, chúng tôi muốn đánh giá khả năng chia sẻ công bằng của MPTFRC so với các giao thức hiện có như: TCP Reno, TFRC tại một đường truyền cổ chai. Mô phỏng gồm hai mô hình mạng như Hình 11.



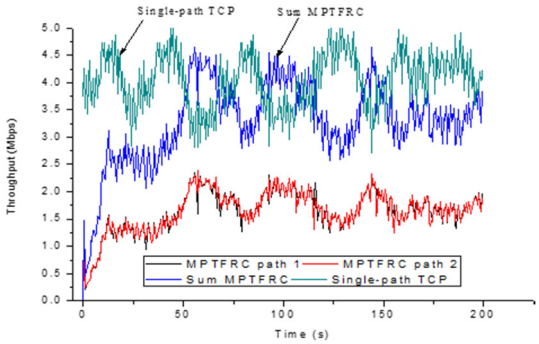
Hình 11: Mô hình mạng mô phỏng chia sẻ công bằng



Hình 12: Thông lượng chia sẻ công bằng giữa MPTFRC với TFRC

Hình 12 cho thấy, mặc dù hai đường MPTFRC cùng cạnh tranh với một luồng TFRC như Hình 11(a), nhưng hai đường MPTFRC không tích cực chiếm thông lượng của đường TFRC. Thông lượng hai đường MPTFRC chỉ bằng gần một nửa so với thông lượng trên đường TFRC. Kết quả này là do α_r đo được trung bình là: $\alpha_1 \approx \alpha_2 \approx 0.25$, nên đưa vào công thức (1) thì tốc độ trên mỗi đường sẽ

bằng 1/2 so với TFRC (vì chúng có chung RTT, băng thông và cùng mức độ tắc nghẽn).

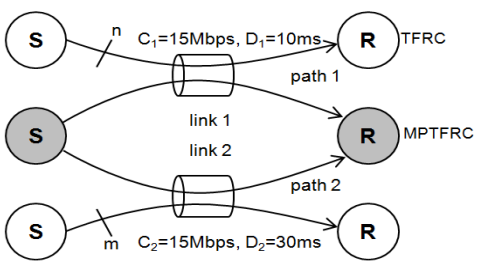


Hình 13: Thông lượng chia sẻ công bằng giữa MPTFRC với TCP Reno

Trường hợp hai đường MPTFRC cũng cạnh tranh công bằng với đường TCP như trong Hình 11(b). Tổng thông lượng của hai đường MPTFRC gần bằng với thông lượng trên đường TFRC đơn đường, cũng như TCP đơn đường (Hình 13) cùng cạnh tranh qua một link. Tức là MPTFRC chia sẻ công bằng với kết nối TFRC và TCP Reno. Bởi vì công thức TFRC được rút ra từ TCP Reno (hay thông lượng TFRC tương đương TCP Reno) nên không khó để giải thích tổng thông lượng MPTFRC hai đường truyền tương đương với thông lượng của TCP Reno trong cùng điều kiện mạng.

c. Đánh giá theo tiêu chí cân bằng tắc nghẽn

Chúng tôi đánh giá tiêu chí cân bằng tắc nghẽn bằng mô hình mạng Hình 14. Mô hình mạng này gồm một kết nối MPTFRC có 2 đường: path 1 và path 2. Trong đó path 1 cạnh tranh với đường TFRC, có n= 20 luồng qua link 1, tốc độ C_1 = 15 Mbps, độ trễ D_1 = 10 ms; path 2 cạnh tranh với một đường TFRC qua link 2 có m= 12 luồng, tốc độ C_2 = 15 Mbps, độ trễ D_2 = 30 ms. Chúng tôi tạo 20 luồng TFRC trên link 1 để mức độ tắc nghẽn cao trên đường thứ 1 (path 1).



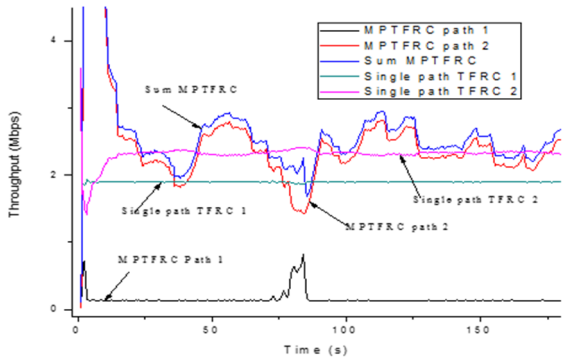
Hình 14: Mô hình mạng mô phỏng cân bằng tắc nghẽn

Tỉ lệ mất gói tin đo được trên hai đường cụ thể như sau:

Bảng 1: Tỉ lệ mất gói tin

Path 1	Path 2
$P_1 = 0.69\%$	$P_2 = 0.12\%$

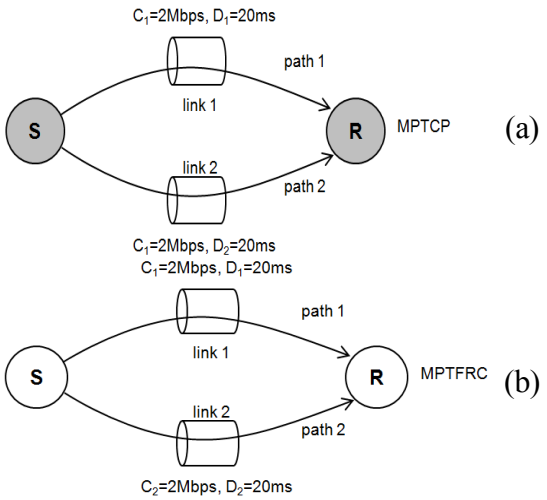
Vì tỉ lệ mất gói tin $p_1 > p_2$, tức là luồng MPTFRC chỉ tập trung truyền thông lượng vào path 2 là chính, để đạt được tiêu chí tăng thông lượng, MPTFRC gửi càng nhiều càng tốt trên path 2 trong khi tốc độ truyền dữ liệu trên path 1 gần như bằng không, nhưng tổng thông lượng của MPTFRC xấp xỉ thông lượng đường tốt nhất trên TFRC (Single-path TFRC2) trong Hình 15.



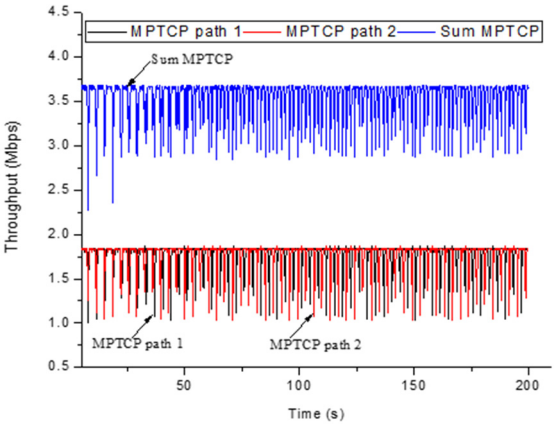
Hình 15: Thông lượng cân bằng tắc nghẽn giữa MPTFRC và TFRC

d. Đánh giá theo tiêu chí sự mượt của tốc độ truyền

Một trong những mục tiêu quan trọng trong thiết kế giao thức cho các ứng dụng multimedia là sự mượt của tốc độ truyền dữ liệu. Đánh giá tốc độ mượt của dữ liệu truyền sử dụng mô hình mạng Hình 16.



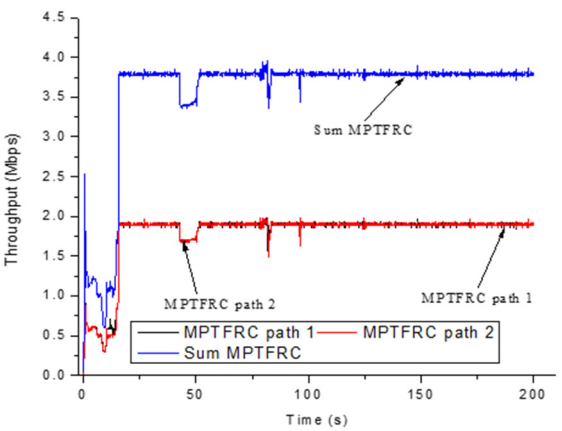
Hình 16: Mô hình mạng mô phỏng sự mượt của tốc độ truyền



Hình 17: Tốc độ dữ liệu truyền mượt của MPTCP

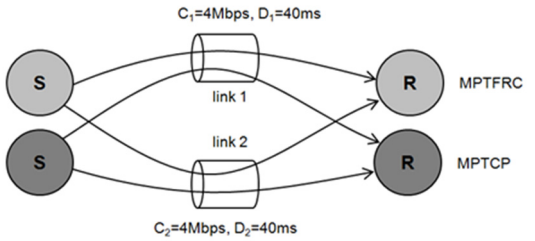
Sự mượt của tốc độ truyền dữ liệu của hai đường MPTCP có sự dao động lớn như Hình 17, tốc độ truyền dữ liệu liên tục thay đổi. Vì trong MPTCP, thuật toán điều chỉnh tốc độ trong giai đoạn giảm tương tự như TCP nên tốc độ có sự dao động lớn.

Ngược lại, sự mượt của tốc độ truyền dữ liệu của hai đường MPTFRC như Hình 18, thông lượng tăng chậm hơn so với MPTCP, tuy nhiên tốc độ truyền trong MPTFRC ít thay đổi, biên độ thay đổi nhỏ. Do MPTFRC thay đổi tốc độ dựa vào công thức (như TFRC).

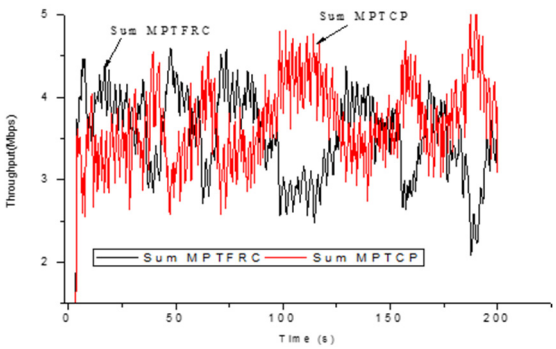


Hình 18: Tốc độ dữ liệu truyền mượt của MPTFRC

Tiếp theo, chúng tôi mô phỏng so sánh tốc độ truyền dữ liệu mượt của hai đường MPTFRC và hai đường MPTCP song song đều qua hai link: link 1 và link 2 có độ trễ D= 40 ms và tốc độ C= 4 Mbps như Hình 19.



Hình 19: Mô hình mạng mô phỏng tốc độ dữ liệu truyền mượt của hai đường MPTFRC và hai đường MPTCP truyền song song



Hình 20: Tốc độ dữ liệu truyền mượt của hai đường MPTFRC và hai đường MPTCP truyền song song

Kết quả Hình 20 cho thấy rằng, tốc độ truyền dữ liệu của hai đường MPTCP và MPTFRC qua hai link song song gần như tương đương. Tuy nhiên, thay đổi tốc độ trên MPTFRC tương đối chậm, tốc độ thay đổi giữa hai lần liên tiếp không lớn. Cụ thể xem xét thời gian từ giây 100 đến giây 150, biên độ dao động tốc độ của MPTCP lớn hơn so với MPTFRC. Đồng thời, qua kết quả mô phỏng cũng cho thấy rằng hai Subflow của hai đường MPTFRC chia sẻ công bằng và cùng tồn tại với hai Subflow của hai đường MPTCP.

4 KẾT LUẬN

Trong bài báo này, chúng tôi đã trình bày kỹ thuật mở rộng điều khiển tắc nghẽn cho ứng dụng multimedia trong giao thức truyền đơn đường vào trong giao thức truyền đa đường. Cải tiến giao thức bao gồm xác định trọng số động nhằm điều chỉnh thông lượng gửi dữ liệu giữa các đường trong nhiều điều kiện mạng khác nhau. Trọng số động của MPTFRC sẽ điều chỉnh lưu lượng tỉ lệ nghịch với mức độ tắc nghẽn của mạng, để dễ dàng đạt được 4 tiêu chí đã nêu bao gồm: *(i) nâng cao thông*

lượng truyền, (ii) đảm bảo chia sẻ công bằng đối với các luồng TCP hiện có; (iii) có khả năng cân bằng tắc nghẽn; (iv) tốc độ dữ liệu truyền mượt. Tuy nhiên, khi cài đặt ở mức gói, thiết kế ban đầu sẽ không cho kết quả mong muốn, đó là sự đánh võng ở mức gói (các đường ít khi sử dụng đồng thời ngay cả khi các đường có cùng điều kiện mạng như băng thông, độ trễ). Để sử dụng đường truyền không bị đánh võng giữa các đường được thực hiện bằng cách tính tham số tránh đánh võng giữa các đường dựa trên đường có kích thước cửa sổ lớn nhất và đường tốt nhất trong quá trình truyền dữ liệu. Mặc dù vậy, nó không làm thay đổi bài toán gốc ban đầu.

Bằng cách mô phỏng bởi nhiều mô hình mạng với các thông số mạng khác nhau, chúng tôi đã đánh giá giao thức đề xuất theo 4 tiêu chí nêu trên. Đồng thời chúng tôi so sánh sự chia sẻ công bằng và cùng tồn tại với các giao thức khác hiện có như TCP Reno, TFRC, MPTCP.

TÀI LIỆU THAM KHẢO

1. Dragana Damjanovic , Michael Welzl, "MulTFRC: Providing Weighted Fairness for Multimedia Applications (and others too!) ", 2009, ACM SIGCOMM Computer Communication Review,NY, USA,volume 39 (number 3), pages 5-12.
2. Damon Wischik, Mark Handley, Costin Raiciu, 2009, "Control of Multipath TCP and Optimization of Multipath Routing in the Internet". NET-COOP 2009, LNCS 5894, pages 204–218.
3. S. Floyd, M. Handley, 2008 "TCP Friendly Rate Control (TFRC): Protocol Specification", Network Working Group, Obsoletes: 3448, Updates: 4342, RFC 5348.
4. Damon Wischik, Costin Raiciu, Adam Greenhalgh, Mark Handley, 2011, "Design, implementation and evaluation of congestion control for multipath TCP" Proceedings of the 8th USENIX conference, CA, USA, pages 8-22.
5. C.Raiciu, M.Handly, D.Wischik, 2011, "Coupled Congestion Control for Multipath Transport Protocols ", Internet Engineering Task Force, RFC 6356, ISSN: 2070-1721.
6. NS-2 network simulator 2, http://www.isi.edu/nsnam/ns/, 26/5/2013.
7. Ramin Khalili,Nicolas Gast, Miroslav Popovic, Utkarsh Upadhyay, Jean-Yves Le Boudec "MPTCP is not Pareto-Optimal: Performance Issues and a Possible Solution", 2012, CoNEXT 12, pages 10-13
8. A. Ford, C.Raiciu, S.Barre, 2011, "Architectural Guidelines for Multipath TCP Development", Internet Engineering Task Force (IETF), RFC 6182 (Informational).
9. Jitendra Padhye, Victor Firoiu, Don Towsley, Jim Kurose, 1998, "Modeling TCP Throughput: ASimple Model and its Empirical Validation", ACM SIGCOMM, vol. 28, pp. 303-314.
10. B. Wang, W. Wei, Z. Guo, and D. Towsley, 2009, "Multipath live streaming via TCP: Scheme, performance and benefits", ACM Trans, NY, USA,vol. 5, no. 3, pages
11. J.-W. Park, R. Karrer, and J. Kim, 2011 "TCP-ROME: A transport-layer parallel streaming protocol for real-time online multimedia environments", Communications and Networks Journal, vol. 13, no. 3, pages. 277 – 285.
12. S. Tullimas, T. Nguyen, R. Edgecomb, and S.-c. Cheung, 2008, "Multimedia streaming using multiple TCP connections," ACM Trans. Multimedia Comput. Commun. Appl., vol. 4, no. 2, pages. 120-121.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

HỆ THỐNG GỢI Ý ÁP DỤNG CHO TRANG WEB TỔNG HỢP TIN TỨC TỰ ĐỘNG

Đỗ Thành Nhân[1] và Trần Nguyễn Minh Thư[2]

[1] *THPT Lê Anh Xuân, tỉnh Bến Tre*
[2] *Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Recommender system for news aggregation website

Từ khóa:
Hệ thống gợi ý, hệ thống hỗ trợ quyết định

Keywords:
Recommender systems, decision support systems

ABSTRACT

To assist the reader faces the information explosion, we built the recommender system applied for a news website automatically (NewsRES). The NewsRES based on the content-based method and collaborative method. The content-based method is used in comparison the content of information or describing news in order to find out the similar news which the users used to be concerned. The CF method passes the tastes of users to take advice or predictions about unknown tastes for other users. The system is applied to 280 students grade 10, 11 at Le Anh Xuan high school for a week. We gain the results: 30.59% of precision, 94.17% of recall and 45.26% of F-measure.

TÓM TẮT

Việc cập nhật tin tức là nhu cầu không thể thiếu trong thời đại hiện nay. Với trang web tổng hợp tin tức, người đọc sẽ gặp một số trở ngại trong việc tìm đọc những thông tin theo ý thích vì sự gia tăng về số lượng cũng như đa dạng về nội dung của tin tức. Nhằm hỗ trợ người đọc đối mặt với sự bùng nổ thông tin, chúng tôi xây dựng hệ thống gợi ý áp dụng cho một trang web tổng hợp tin tức tự động (NewsRES). NewsRES sử dụng phương pháp lọc theo nội dung (content-based) được thực hiện dựa trên việc so sánh nội dung thông tin hay mô tả tin tức để tìm ra những tin tức tương tự với những gì mà người dùng đã từng quan tâm; phương pháp phối hợp (CF) thông qua các thị hiếu đã được biết đến của một nhóm người dùng để đưa các tư vấn hoặc dự đoán về thị hiếu chưa biết cho một số người dùng khác. Hệ thống này được áp dụng cho 280 học sinh lớp 10, 11 tại trường trung học Lê Anh Xuân, Bến Tre. Kết quả thực nghiệm trên hệ thống NewsRES: Precision 30.59%, Recall 94.17% và F-Measure 45.26%.

1 GIỚI THIỆU

Trong những năm gần đây, hệ thống gợi ý (recommender system) được biết đến như là một sự phát triển quan trọng trong việc giúp người dùng đối mặt với sự bùng nổ thông tin. Hệ thống này được ứng dụng trong nhiều lĩnh vực như thương mại điện tử với Amazon [4], Netflix [12], Ebay [10]; trong lĩnh vực giải trí với MovieLens, Last.fm, Film-Conseil [6]; trong lĩnh vực khác như tin tức trực tuyến netnews [7],…

Kể từ năm 2007, đã có hội nghị chuyên về hệ gợi ý (ACM) là diễn đàn quốc tế hàng đầu cho việc trình bày kết quả nghiên cứu mới, trong lĩnh vực rộng lớn của hệ gợi ý.

Tuy nhiên, các hệ thống gợi ý hiện tại vẫn đòi hỏi phải có nhiều cải tiến hơn nữa để làm cho

phương pháp gợi ý hiệu quả hơn phù hợp với từng lĩnh vực (loại dữ liệu) áp dụng để có thể cung cấp gợi ý phù hợp với từng cá nhân riêng biệt [3], [9].

Trong khuôn khổ nghiên cứu này, chúng tôi muốn hướng tới hệ thống gợi ý áp dụng cho một trang web tổng hợp tin tức tự động. Với trang web tổng hợp tin tức, người đọc sẽ gặp một số trở ngại trong việc tìm đọc những thông tin theo ý thích vì sự gia tăng về số lượng cũng như đa dạng về nội dung của tin tức. Sự ra đời kỹ thuật Really Simple Syndication (RSS)[7] và sự phong phú về số lượng các trang báo điện tử hiện nay là cơ sở để chúng tôi xây dựng một trang tổng hợp tin tức tự động. Trang web này ra đời còn nhằm mục đích tránh bất tiện cho người dùng trong việc phải mở nhiều trang tin tức khác nhau.

2 HỆ THỐNG NEWSRES

Có rất nhiều cách để dự đoán, ước lượng hạng/điểm cho các dữ liệu như sử dụng học máy, lí thuyết xấp xỉ, các thuật toán dựa trên kinh nghiệm… Các hệ thống gợi ý thường được phân thành ba loại dựa trên cách nó dùng để ước lượng các đánh giá về sản phẩm:

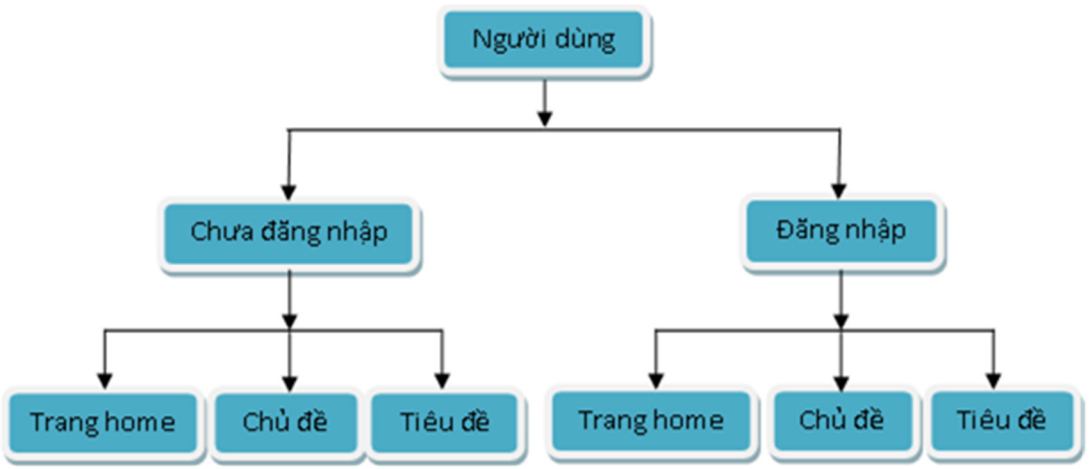
– **Dựa trên nội dung (content-based)**[1]: người sử dụng ðược gợi ý **mục dữ liệu (item)** týõng tự như những mục dữ liệu ðược người sử dụng thích trong quá khứ.

– **Gợi ý phối hợp (collaborative filtering)**[1]: người sử dụng ðược gợi ý mục dữ liệu của những người có cùng **"khẩu vị" và "sở thích"** với mình.

– **Gợi ý hỗn hợp (hybrid)**[1]: kết hợp cả hai tiếp cận ở trên.

2.1 Mô tả hệ thống NewsRES

Hệ thống NEWSRES xây dựng cho hai trường hợp: khi người dùng đăng nhập vào hệ thống hoặc không đăng nhập vào hệ thống như lưu đồ (Hình 1).



Hình 1: Lưu đồ tổng quát

2.1.1 Dữ liệu đầu vào của hệ thống NewsRES

Phân tích dữ liệu của hệ thống sẽ xây dựng để xác định giải thuật sẽ sử dụng. Dữ liệu đầu vào của hệ thống lấy được từ công nghệ RSS ta được:

- Tiêu đề.
- Phân loại/ nhóm tin.
- Tóm tắt.
- Nội dung.
- Ngày tháng.

Lịch sử truy cập của người dùng: Khi người dùng đăng ký thông tin, hệ thống sẽ lưu lại những thông tin người dùng (như nghề nghiệp, sở thích, quan tâm,…). Ngoài ra hệ thống lưu lại lịch sử truy cập tin tức của người dùng như:

– Người dùng đọc thể loại nào bao nhiêu lần trong khoảng thời gian k;

– Người dùng đọc tin "a" rồi đọc tiếp những tin nào.

2.1.2 Đặc trưng của hệ thống NewsRES

Hệ thống gợi ý tin tức là một lĩnh vực giàu tiềm năng bởi số lượng các sản phẩm tư vấn, số lượng người dùng và số lượt sử dụng tương đối nhiều. Tuy nhiên, đi kèm theo đó là các thử thách về các đặc trưng riêng của miền đối tượng tin tức cũng như các đặc trưng chung của người sử dụng gợi ý.

Tin tức là một đối tượng gợi ý đặc biệt, các đặc trưng[10] sau của tin tức giúp đưa ra các giải thuật hữu hiệu hơn trong xây dựng giải thuật cho hệ thống gợi ý tin tức của đề tài:

– Tính thời gian: theo thời gian, tin tức mất đi giá trị. Hệ thống NewsRES gợi ý từ dữ liệu được lấy trong khoảng thời gian 'x' ngày.

– Tính đa quan tâm: tại một thời điểm, người dùng có thể có nhiều mối quan tâm khác nhau. Hệ thống gợi ý phải cung cấp cho người đọc tin tức theo nhiều loại chủ đề chứ không chỉ gợi ý các tin của duy nhất một chủ đề. Ví dụ: họ có thể quan tâm đến cả các thông tin về cả thể thao và chính trị.

2.2 Giải thuật

Hệ thống gợi ý tin tức NewsRES được xây dựng tập trung vào hai giải thuật: giải thuật gợi ý **dựa trên nội dung và** giải thuật **gợi ý phối hợp**.

2.2.1 Áp dụng giải thuật gợi ý dựa trên nội dung (TF-IDF)

Phương pháp lọc theo nội dung được thực hiện dựa trên việc so sánh nội dung thông tin hay mô tả tin tức để tìm ra những tin tức tương tự với những gì mà người dùng đã từng quan tâm để giới thiệu cho người dùng những tin tức này. Lọc dựa trên nội dung thực hiện hiệu quả trên các đối tượng dữ liệu biểu diễn dưới dạng văn bản.

Lọc dựa trên nội dung không gặp phải các vấn đề rất khó giải quyết của lọc cộng tác trên miền đối tượng tin tức: Các tin tức liên tục được sinh ra và cần dễ dàng tiếp cận trong khi quá trình lọc cộng tác không thể tạo ra các sản phẩm chưa từng được đánh giá bởi người dùng khác hoặc những người dùng chưa từng đánh giá một sản phẩm nào. Khó tìm ra được các sản phẩm đã được đánh giá bởi một lượng đủ người dùng vì số lượng quá lớn các tin tức mới và đặt gánh nặng cung cấp thông tin đánh giá lên người dùng.

Vì những lý do trên, hệ thống có áp dụng giải thuật gợi ý dựa trên nội dung để gợi ý một số tin cho người đọc.

Giải thuật dựa trên nội dung

Đầu vào:

– Tập M chứa danh sách các tiêu đề (document).

– Tập Q chứa tiêu đề cơ sở (tiêu đề cần gợi ý).

Đầu ra:

– Tập C: danh sách các tin tức được xếp hạng dựa theo độ ưu tiên gợi ý.

Giải thuật:

B1: Xử lý dữ liệu (Tập M và Q):

– Đưa về chữ thường.
– Loại bỏ từ dừng (stop word).
– Loại bỏ kí tự đặc biệt.
– Loại bỏ chữ số.

B2: Tính D_f & IDF

Ta có công thức:

$$IDF(w)=\log(N/D_f(w))$$

Trong đó:

– N là tổng số lượng tài liệu cần tý vấn cho người sử dụng

– Df(w) là số lượng tài liệu mà một từ nào đó xuất hiện

– w là 1 từ nào đó.

B3: Tính trọng số TF & IDF

Ta có công thức:

$$TF= t_f / f$$

$$W= TF*IDF$$

Trong đó:

– t_f: Số lần xuất hiện của từ t trong tài liệu f.
– f: Tổng số các từ trong tài liệu f.
– W: Trọng số.

B4: Tính Normalizing Vectors

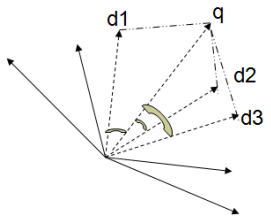
Tìm hiểu mô hình Vector Space Model (VSM):

Vector trong không gian 2 chiều thể hiện là ax+by. Tương tự với không gian n chiều. Mỗi vector là một danh sách các hệ số [a,b] định nghĩa độ lớn của vector trong chiều đó. Mỗi từ trong câu truy vấn là một chiều trong VSM, nếu câu truy vấn có 'n' từ → là một vector n-chiều. Mỗi một tài liệu cũng là một vector nhiều chiều. Như vậy, tiêu đề tin tức cần truy vấn và tiêu đề trong cơ sở dữ liệu là những vector nhiều chiều. Ta cần tính điểm (Score) giữa tiêu đề câu tư vấn và tiêu đề trong cơ sở dữ liệu.

Có hai cách tính Score $\{\vec{q},\vec{d}\}$

– Sử dụng góc giữa $\vec{q},\vec{d}$

– Sử dụng khoảng cách $\vec{q},\vec{d}$



Hình 2 : Hình vector câu truy vấn

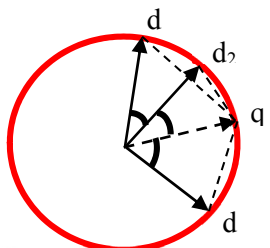
Trong đó:

– $\vec{q}$: tiêu đề tin tức cần tư vấn

– $\vec{d}$: tiêu đề trong cơ sở dữ liệu.

Mỗi tài liệu có độ dài khác nhau thì cách tính theo khoảng cách không còn đúng nữa vì tài liệu nào càng dài thì score càng lớn. Từ đó ta cần **Normalizing Vectors**, làm cho các vector có cùng độ lớn.

Công thức:



$$Q' = \left(\frac{q_1}{\sqrt{\sum_{i=1}^{m} q_i^2}}, \frac{q_2}{\sqrt{\sum_{i=1}^{m} q_i^2}}, \ldots, \frac{q_m}{\sqrt{\sum_{i=1}^{m} q_i^2}}\right)$$

$$d' = \left(\frac{d_1}{\sqrt{\sum_{i=1}^{m} d_i^2}}, \frac{d_2}{\sqrt{\sum_{i=1}^{m} d_i^2}}, \ldots, \frac{d_m}{\sqrt{\sum_{i=1}^{m} d_i^2}}\right)$$

Trong đó: q, d: là trọng số TF*IDF

B5: Tính độ tương đồng của chúng bằng độ đo cosin

$$u(c,s) = \frac{\sum_{i=1}^{K} w_{ic} w_{is}}{\sqrt{\sum_{i=1}^{K} w_{ic}^2} \sqrt{\sum_{i=1}^{K} w_{is}^2}}$$

2.2.2 Áp dụng giải thuật gợi ý phối hợp (CF)

Phương pháp lọc phối hợp được thực hiện thông qua thị hiếu đã được biết đến của một nhóm người dùng để đưa các tư vấn hoặc dự đoán về thị hiếu chưa biết cho một số người dùng khác. Lọc phối hợp sử dụng cơ sở dữ liệu về sở thích của người dùng đối với các item để dự đoán các chủ đề hoặc sản phẩm thêm vào cho một người dùng mới có cùng sở thích.

Hệ thống gợi ý cộng tác khắc phục được nhiều nhược điểm của hệ thống dựa trên nội dung. Một điểm quan trọng là nó có thể xử lý mọi loại dữ liệu và gợi ý một loại sản phẩm, kể cả những sản phẩm mới, khác hoàn toàn so với những gì người dùng từng xem.

Vì những lý do trên, thay vì chỉ cần dùng giải thuật gợi ý đựa trên nội dung, hệ thống đề xuất thêm giải thuật gợi ý phối hợp dựa trên bộ nhớ.

Giải thuật độ tương quan Pearson giữa hai item (CF-ITEM):

Đầu vào:

– Cho tập người dùng u thuộc U là những người cùng đánh giá về hai item i và j

– **Đầu ra:**

– Độ tương quan Pearson giữa item i và item j.

Giải thuật:

B1: Tính trung bình của item thứ I bởi những người dùng khác

B2: Tính Độ tương quan Pearson giữa item i và item j

Công thức:

$$w_{i,j} = \frac{\sum_{u \in U}(r_{u,i} - \overline{r_i})(r_{u,j} - \overline{r_j})}{\sqrt{\sum_{u \in U}(r_{u,i} - \overline{r_i})^2} \sqrt{\sum_{u \in U}(r_{u,j} - \overline{r_j})^2}}$$

Trong đó:

– $r_{u,i}$: là đánh giá của người dùng u cho item i,

– $\overline{r_i}$: là đánh giá trung bình của item thứ I bởi những người dùng khác.

– Đánh giá người dùng u cho item i: số lần click chuột trên mục tin.

Giải thuật độ tương quan Pearson giữa người dùng (CF-USER):

Đầu vào:

– Cho tập người dùng u thuộc U

Đầu ra:

– Độ tương quan Pearson giữa user U_i và user U_j.

Giải thuật:

B1: Tính trung bình của người dùng U

B2: Tính Độ tương quan Pearson giữa user U_i và user U_j

Công thức:

$$w_{ij} = \frac{\sum_{x \in P_i \cap P_j}(r_{i,x} - \overline{r_i})(r_{j,x} - \overline{r_j})}{\sqrt{\sum_{x \in P_i \cap P_j}(r_{i,x} - \overline{r_i})^2}\sqrt{\sum_{x \in P_i \cap P_j}(r_{j,x} - \overline{r_j})^2}}$$

Trong đó:

– $x \in P_i \cap P_j$: là tập sản phẩm mà người dùng *i* và người dùng *j* cùng đánh giá

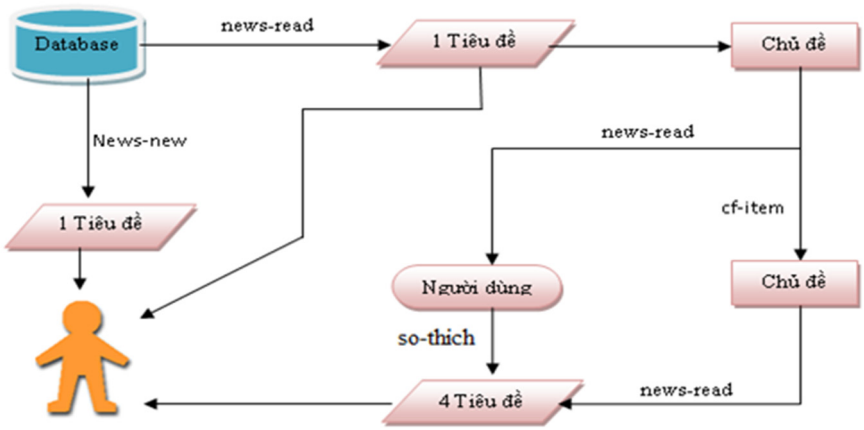
– $r_{i,x}$: là đánh giá của người dùng *i* lên sản phẩm *x*.

– $\overline{r_i}$: là đánh giá trung bình của người dùng *i*.

2.3 Giới thiệu hệ thống NewsRec

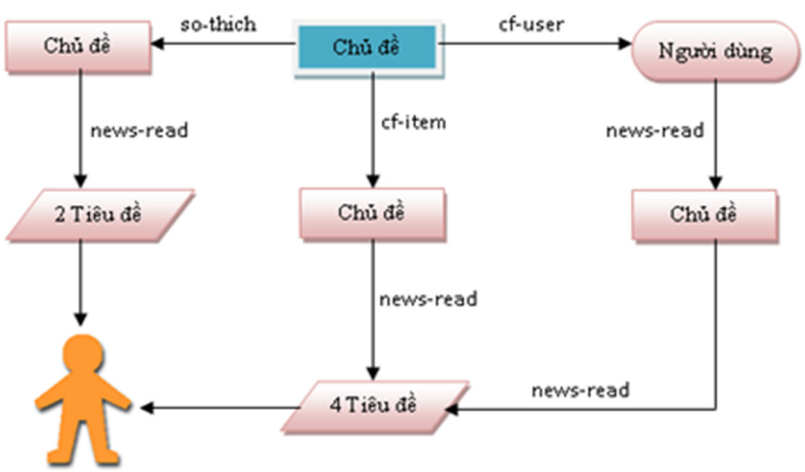
Hệ thống dự đoán thông qua danh sách Top-N tin tức được sắp xếp theo thứ tự giảm dần về độ tương quan. Trong khuôn khổ bài báo này, chúng tôi trình bày 2 trạng thái của người dùng như sau:

Hệ thống gợi ý khi người dùng không đăng nhập: gợi ý một tiêu đề mới nhất trong dữ liệu; một tiêu đề đọc nhiều nhất; lấy chủ đề có tiêu đề đọc nhiều nhất kế tiếp tìm người dùng đọc chủ đề này nhiều nhất dùng giải thuật "so-thich" với người dùng này để gợi ý hai tiêu đề; lấy chủ đề có tiêu đề đọc nhiều nhất dùng giải thuật "cf-item" rồi gợi ý hai tin đọc nhiều nhất như lưu đồ (Hình 3).



Hình 3: Lưu đồ đang ở trang HOME không đăng nhập

Hệ thống gợi ý khi người dùng đăng nhập click vào chủ đề: dùng giải thuật "so-thich" tìm chủ đề gợi ý hai tiêu đề đọc nhiều nhất trong chủ đề này; dùng giải thuật "cf-item" tìm chủ đề tương tự như chủ đề vừa click => gợi ý hai tiêu đề đọc nhiều nhất trong chủ đề,; dùng giải thuật "cf-user" tìm người dùng tương tự với người dùng hiện tại, lấy chủ đề có tiêu đề mà người dùng này đọc nhiều nhất và gợi ý hai tin đọc nhiều nhất trong chủ đề này như lưu đồ (Hình 4).



Hình 4: Lưu đồ click vào chủ đề khi đăng nhập

Chú thích:

– news-new: Những tin mới đăng.

– news-read: Những tin đọc nhiều nhất, trong khoảng thời gian 3 ngày.

– so-thich: truy vấn dựa vào hồ sơ người dùng.

– news-assess: Người dùng đánh giá cao nhất.

– cf-item: Giải thuật độ tương quan Pearson giữa hai item.

cf-user: Giải thuật độ tương quan Pearson giữa người dùng.

2.4 Phương pháp đánh giá

Đề tài trình bày tóm tắt các phương pháp đánh giá hiệu quả hệ thống gợi ý. Đây là một vấn đề rất quan trọng giúp cho người sử dụng có thể lựa chọn đúng mô hình phù hợp với dữ liệu hay ứng dụng trong thực tế.

2.4.1 Nghi thức kiểm tra

Để đánh giá chất lượng của một hệ thống gợi ý chúng ta cần phải đúng cách phân vùng các tập dữ liệu vào một tập huấn luyện và một bộ kiểm tra. Điều rất quan trọng là hiệu suất được tính toán trên dữ liệu mà không có phần trong việc xây dựng các mô hình. Một số chương trình học tập cũng cần một tập hợp xác nhận để tối ưu hóa các thông số mô hình. Bộ dữ liệu thường được chia theo một trong các phương pháp sau:

Holdout: chia tách tập dữ liệu thành hai phần: một tập huấn luyện và một bộ kiểm tra. Những bộ có thể có tỷ lệ khác nhau. Lấy ngẫu nhiên 2/3 tập dữ liệu D để huấn luyện và 1/3 tập dữ liệu còn lại dùng cho bộ kiểm tra, có thể lặp lại quá trình này k lần rồi tính giá trị trung bình.

k-fold: chia tập dữ liệu D thành k phần (fold) bằng nhau, lặp lại k lần, mỗi lần sử dụng k-1 folds để học và 1 fold để kiểm tra, sau đó tính trung bình của k lần kiểm tra. Khi tập dữ liệu D có hơn 300 phần tử, phương pháp thường sử dụng là 10 fold (k=10). Nếu tập D có ít hơn thì leave-1-out (k= số phần tử) được đề nghị sử dụng.

2.4.2 Các chỉ số sử dụng đánh giá

Việc đánh giá hiệu quả của một hệ thống khuyến nghị là cần thiết. Tuy nhiên, việc đánh giá không chỉ phụ thuộc vào dữ liệu mà còn phụ thuộc vào mục tiêu của hệ thống khuyến nghị [2]. Thật vậy, một số hệ thống nhấn mạnh sự đa dạng của các mục trong danh sách khuyến nghị, trong khi những người khác tập trung vào tính mới. Có nhiều chỉ tiêu đánh giá khác nhau nhưng trong khuôn khổ bài báo này, chúng tôi sử dụng precision, recall và F- Measure các chỉ số chủ yếu được sử dụng trong các hệ thống khuyến nghị của thương mại điện tử[10, 11].

$$precision = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{tp + fn}$$

$$F - measure = \frac{2\ x\ precision\ x\ recall}{precision + recall}$$

Ở đó:

– true positive (TP): tin tức thú vị gợi ý cho người dùng),

– true negative (TN): tin tức thú vị không gợi ý cho người dùng),

– false negative (FN): tin tức không thú vị không gợi ý cho người dùng),

– false positive (FP) : tin tức không thú vị gợi ý cho người dùng).

3 KẾT QUẢ VÀ THẢO LUẬN

3.1 Xây dựng tập dữ liệu

Kết quả nghiên cứu này được thực nghiệm với tập dữ liệu tin tức (bộ dữ liệu NewsRES). Dữ liệu này được lấy tin tự động từ hai trang web tin tức (vnExpress.net, dantri.com.vn). Ngoày ra, thông tin người dùng được lưu lại từ thông tin đăng ký sử dụng và nhật ký sử dụng của người dùng. Đây là dữ liệu đầu vào của hệ thống đã được mô tả trong mục 2.1.1. Tất cả các dữ liệu này được dùng để xây dựng hệ thống gợi ý áp dụng cho trang web tổng hợp tin tức tự động.

Hệ thống này bước đầu được áp dụng cho học sinh trường THPT Lê Anh Xuân, Bến Tre. Thực nghiệm trên bốn lớp khối 10, 11 (10a, 10c1, 11a, 11c1, 11c2, 11c4, 11c5). NewsRES tính thời điểm thực nghiệm có tổng số 1020 tin và 280 người dùng, có được 229 giao dịch (session), số tin gợi ý cho người dùng (Recommendhistory) 6481 tin, tổng số tin người dùng đọc 1976 tin.

3.2 Phương pháp thử nghiệm

Trong trường hợp phân lớp nhị phân (lớp gợi ý là lớp cần quan tâm, lớp không gợi ý). Trong phân lớp nhị phân, người ta cần tính ma trận phân lớp C có kích thước 2x2 (Bảng 1).

Bảng 1: Bảng ma trận phân lớp C

dự đoán =>	Gợi ý	Không gợi ý
Dùng	True – Positive (tp)	False-Negative (fn)
Không dùng	False – Positive (fp)	True – Negative (tn)

Trong đó:

– True positive (tp): Số tin tức tư vấn chính xác.

– False negative (fn): Số tin tức dùng mà không có tư vấn.

– False positive (fp): Số tin tức tư vấn không chính xác.

– True negative (tn): Số tin tức không dùng không tư vấn.

Các độ đo được tính thông qua precision, recall và F- Measure xác định theo công thức. Giá trị precision, recall càng lớn hiệu quả phương pháp càng cao[2].

$$precision = \frac{Số\ tin\ tức\ tư\ vấn\ chính\ xác}{Tổng\ số\ tin\ tức\ tư\ vấn} = \frac{tp}{tp + fp}$$

$$recall = \frac{Số\ tin\ tức\ tư\ vấn\ chính\ xác}{Tổng\ số\ tin\ tức\ trong\ tập\ người\ dùng} = \frac{tp}{tp + fn}$$

$$F - measure = \frac{2\ x\ precision\ x\ recall}{precision + recall}$$

Ví dụ chúng ta có thể xét giao dịch id là 7e02fa0676, tập dữ liệu có 940 tin trong đó có 15 tin sử dụng là lớp người dùng quan tâm và 913 tin thuộc lớp không dùng. Ta tính được các sai số dự đoán sau:

Bảng 2: Ma trận phân lớp

dự đoán =>	Gợi ý	Không gợi ý
Dùng	12	3
Không dùng	28	897

$$precision = \frac{Số\ tin\ tức\ tư\ vấn\ chính\ xác}{Tổng\ số\ tin\ tức\ tư\ vấn} = \frac{12}{12 + 28} = 30\%$$

$$recall = \frac{Số\ tin\ tức\ tư\ vấn\ chính\ xác}{Tổng\ số\ tin\ tức\ trong\ tập\ người\ dùng} = 80\%$$

$$F - measure = \frac{2\ x\ precision\ x\ recall}{precision + recall} = \frac{2\ x\ 30\%\ x\ 80\%}{30\% + 80\%} = 43.64\%$$

3.3 Kết quả thử nghiệm

Hệ thống NewsRES là mô hình gợi ý tin tức. Độ precision, recall và F-Measure được tính toán dựa trên 1020 tin dùng để tư vấn. Thông qua kiểm nghiệm, chúng tôi được kết quả kiểm nghiệm của mô hình đề xuất trong trường hợp khi người dùng đăng nhập và người dùng không đăng nhập được thể hiện trong hai bảng (Bảng 3, Bảng 4).

Bảng 3: Kết quả thực nghiệm khi người dùng đăng nhập

SessionID	Precision	Recall	F-Measure
084a7a6be3	26.73%	87.10%	40.91%
1663e275a7	36.36%	98.11%	53.06%
4998df2c47	18.75%	100%	31.58%
398ca839ae	25.23%	93.10%	39.70%
24878b0e96	30.53%	100%	46.78%
6afde6425b	50%	100%	66.67%
90b50df2dc	16.67%	100%	28.58%
7e36e77535	23.88%	100%	38.55%
…	…	…	…
36670fc3b6	25%	100%	40%
8dac844637	21.62%	100%	35.55%
8dbc531d17	16.67%	80%	27.59%
5d5c751399	10%	50%	16.67%
27d8e225a4	50%	100%	66.67%
518ee45dc7	50%	100%	66.67%
7e02fa0676	35.71%	83.33%	50%
fbb077a2d0	25%	100%	40%
2486dac3a9	44.44%	100%	61.53%
Tổng	30.59%	94.17%	45.26%

Bảng 4: Kết quả thực nghiệm khi người dùng không đăng nhập

SessionID	Precision	Recall	F-Measure
672991a038	42.86%	100%	60%
24878b0e96	34.48%	100%	51.28%
9f0f1540b2	20%	50%	28.57%
7ed88d577c	33.72%	100%	50.43%
4998df2c47	27.27%	100%	42.85%
1663e275a7	14.29%	100%	25.01%
613b3746d5	40%	100%	57.14%
9b79d29215	20.22%	100%	33.64%
8c1e09d46d	10.34%	75%	18.17%
…	…	…	…
6afde6425b	7.14%	100%	13.33%
24eb7fbfc3	36.36%	100%	53.33%
7e02fa0676	26.92%	77.78%	40%
5a212af2c6	40%	100%	57.14%
4738ac737a	22.22%	100%	36.36%
57102845ad	12.50%	100%	22.22%
ed42411619	22.86%	100%	37.21%
Tổng	25.13%	86%	38%

Kết quả thực nghiệm cho thấy (Precision 30.59%, Recall 94.17% và F-Measure 45.26%.) thì hệ thống NewsRES hiệu quả hơn khi người dùng

không đăng nhập vào hệ thống (Precision 25.13%, Recall 86% và F-Measure 38%).

4 KẾT LUẬN VÀ ĐỀ XUẤT

4.1 Kết luận

Chúng tôi đã trình bày mô hình hệ thống gợi ý áp dụng cho trang web tổng hợp tin tức tự động và hiệu quả kết hợp lọc nội dung và lọc cộng tác để gợi ý tin tức cho người dùng.

Đề tài đã tiến hành thử nghiệm mô hình trên tập dữ liệu (NewsRES) có 940 tin được lấy tự động từ hai trang web vnExpress.net, dantri.com.vn, người sử dụng là học sinh trường THPT Lê Anh Xuân khối 10, 11. Kết quả kiểm nghiệm trên tập dữ liệu NewsRES, ta có kết quả với Precision = 30.59%, Recall = 94.17% , F-Measure = 45.26%.

Theo công trình nghiên cứu, phát triển và ứng dụng CNTT-TT, Lọc cộng tác và lọc theo nội dung dựa trên mô hình đồ thị, năm 2009 của Nguyễn Duy Phương, Từ Minh Phương[5] thì độ đo Precision = 29.2%. Tuy không thể so sánh trực tiếp kết quả thực nghiệm của chúng tôi so với công trình trong bài báo [5], nhưng kết quả này cũng phản ảnh được các hệ thống gợi ý hiện tại chưa đạt được giá trị precision cao như những lĩnh vực nghiên cứu khác.

4.2 Đề xuất

Tiến hành thử nghiệm hệ thống NewsRES với nhiều đối tượng khác nhau (giáo viên, học sinh,…) với khoảng thời gian nhiều hơn.

Tìm kiếm các dữ liệu trong cùng lĩnh vực để so sánh, đối chiếu kết quả nghiên cứu với những giải pháp khác.

Hoàn thiện hệ thống gợi ý người đọc cho trang web tổng hợp tin tức tự động thông qua việc đánh giá kết quả gợi ý và phản hồi của người đọc cũng như trong lúc so sánh với các giải pháp khác để tăng chất lượng của các gợi ý.

Phát triển trên các lĩnh vực khác như tìm kiếm khách sạn, địa điểm du lịch.

TÀI LIỆU THAM KHẢO

1. Gendiminas Adomavicius, Alexander Tuzhilin, Toward the Next Generation of Recommender Systems: A Survey of the State-of-the Art and Possible Extensions.
2. Herlocker Jonathan L., Konstan Jo- seph A., “Evaluating collaborative filtering recommender systems” ACM *Trans. Inf. Syst.*, vol. 22, no. 1, pp. 5–53, 2004.
3. Huang, Z.; Zeng, D. & Chen, A comparative study of recommendation algorithms for e-commerce applications, IEEE Intelligent Systems, 2006.
4. Linden, G.; Smith, B. & York, J. , Amazon.com Recommendations: Item-to-Item Collaborative Filtering, IEEE Internet Computing, IEEE Educational Activities Department, 2003, 7, 76-80.
5. Nguyễn Duy Phương, Từ Minh Phương, 2009, Các công trình nghiên cứu, phát triển và ứng dụng CNTT-TT, Lọc cộng tác và lọc theo nội dung dựa trên mô hình đồ thị.
6. Perny, P. & Zucker, J. D., Preference-based Search and Machine Learning for Collaborative Filtering: the “Film-Conseil” recommender system, Information, Interaction, Intelligence, 2001, 1, 9-48.
7. P. Resnick, N. Iacovou, M. Sushak, P. Bergstrom, and J. Riedl. "GroupLens: An Open Architecture for Collaborative Filtering of Netnews”, Proceedings of the 1994 Computer Supported Cooperative Work Conference, ACM, 1994.
8. RSS, http://en.wikipedia.org/wiki/RSS.
9. Sarwar, B. & al., Analysis of recommendation algorithms for e-commerce EC '00, ACM, 2000, 158-167.
10. Schafer, J. B.; Konstan, J. A. & Riedl, J., E-Commerce Recommendation Applications, Data Min. Knowl. Discov., Kluwer Academic Publishers, 2001, 5, 115-153.
11. Uông Huy Long, 2010, khóa luận tốt nghiệp đại học, giải pháp mở rộng thông tin ngữ cảnh phiên duyệt web người dùng nhằm nâng cao chất lượng tư vấn trong hệ thống tư vấn tin tức.
12. Yehuda Koren, August 2009, The BellKor Solution to the Netflix Grand Prize.

Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

RELATIONAL NETWORK OF GEOMETRIC OBJECTS AND APPLY TO DRAW THE FIGURE OF PROBLEM AUTOMATICALLY

Nguyễn Đình Hiển[1] và Phạm Thi Vương[1]
[1] *Trường Đại học Công nghệ thông tin, ĐHQG-HCM*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Mạng quan hệ giữa các đối tượng hình học và ứng dụng giải quyết bài toán vẽ hình tự động

Từ khóa:
Trí tuệ nhân tạo,biểu diễn tri thức, cơ sở tri thức, mạng quan hệ

Keywords:
Artificial intelligence, knowledge representation, knowledge base system, relational network

ABSTRACT

Knowledge base is a central part of intelligent system, so knowledge representation plays an important role in constructing these systems. Nowadays, there are many various methods for representing knowledge. In fact, a popular form of knowledge domain is knowledge about relations, however, the current models are not efficient for executing and reasoning on this kind of knowledge domain. Besides that, automatic drawing geometric objects is an important and basic tool in solving geometry problems. In this paper, we present a method for modelizing the relations between geometric objects. Base on this model, we also introduce an application of this drawing method for solving plane geometry problems in middle school program.

TÓM TẮT

Cơ sở tri thức là thành phần trung tâm của một hệ thống thông minh, vì vậy biểu diễn tri thức đóng một vai trò rất quan trọng trong việc xây dựng các hệ thống thông minh. Tri thức của con người rất đa dạng và phức tạp, trong đó, tri thức quan hệ là một dạng tri thức phổ biến. Tuy nhiên, các mô hình biểu diễn tri thức hiện tại không thật sự hiệu quả cho việc biểu diễn và suy luận cho miền tri thức này trên máy tính. Bên cạnh đó, việc vẽ hình cho các bài toán hình học phẳng là việc rất cần thiết trong quá trình giải toán. Trong bài báo này, chúng tôi sẽ trình bày một phương pháp mô hình hóa tri thức về các quan hệ giữa các đối tượng trong kiến thức hình học phẳng. Trên cơ sở đó, chúng tôi cũng sẽ trình bày một ứng dụng của phương pháp này trong việc thiết kế chương trình vẽ hình tự động cho các bài toán hình học phẳng ở cấp trung học cơ sở.

1 INTRODUCTION

In geometric problems, the figures play a vital and essential role. It help the learner easier to imagine and grasp the important information from the problem. Moreover, the figure and the facts in hypothesis of problem have a very close relationship. The hypothesis can generate a corresponding figure, and from a figure, the information of the hypothesis can be specified adequately. Generally, the figure which perform structures and realions beetween geometric objects is a modeling step for the problem so that the learner can realize visually and quickly the facts from the hypothesis.

For the problems which require drawing, researching and proposing a method for automatic drawing are absolutely necessary and naturally. In books [1,2,8], they have many kinds to perform the realtion beween objects such as: semantic network, conceptual graph, neural network, etc., and the

other relational knowledge model has been proposed in [5]. However, these model cannot to represent geometric relations between objects, especially relations for drawing the figure of problem. In this paper, we are presenting a method for automatic drawing geometry problems of middle school. This approach was carried out based on a logic pattern in terms of semantic relation between objects. It use the idea of semantic net model (see [9, 10]) to form the relational network of geometric objects. Therefore, the program can provide drawing steps in detail and suitably like a human's way of thinking. Meanwhile, for other common drawing methods, it is almost impossible to imitate drawing steps. Furthermore, the program was set up by MAPLE and C#. This is really a useful improvement for the system which supports solving geometry problems in middle school.

2 KNOWLEDGE MODEL ABOUT RELATIONS BETWEEN GEOMETRIC OBJECTS

2.1 Relation Network of Geometric objects

In paper [12], the authors present an Intelligent Problem System to solve plane geometry problems in middle shool. However, this system do not have figrue to illustrate the problem. Base on Computational Object model is presented in [4], combine conceptual graph model [11], we propose an relational network of geometric objects.

This relational network has two components:

(Vertex, Arc)

Each geometry object in this network is an Computation Object. They thay can be divided into two kinds as follows:

* *Deductive kind*: includes objects which will be determined automatically if and only if their components are defined.

Example 2.1:

– Object ANGLE is determined if and only if its two rays are defined.

– Object TRIANGLE is determined if and only if its three sides are defined.

* *Primitive kind:* includes objects that are not deductive kind.

Example 2.2:

– Object POINT.

– Object LINE.

In order to solve the problem of automatic drawing, it is very important and necessary to perform the relationship between Geometric objects. This relation can be represented by the semantic net model which has structure as follow:

Vertex which represents Geometric objects includes two kinds:

– Vertex represents Objects of deductive kind.

– Vertex represents Objects of primitive kind.

The structure of a Vertex includes:

– NAME: the name of the object

– TYPE : the type of the object

– KIND : the kind of the object

– PREVIOUS : the set of previous objects

– F : the set of facts that define the object (facts has been defined in section 2.2)

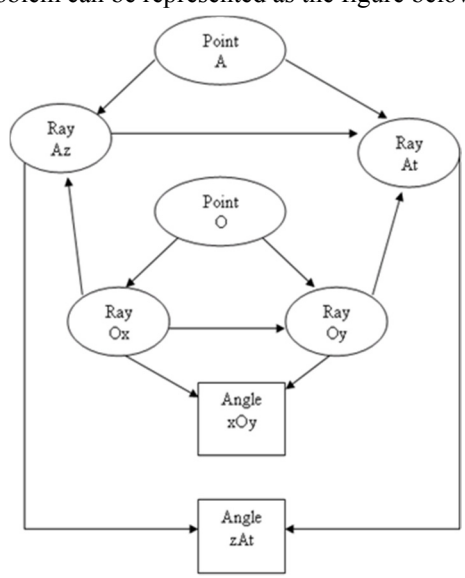
Arc: is used to connect the vertices. It represents the generating relation between objects.

Example 2.3:

Let angle xOy be acute.

Draw acute angle zAt such that Az // Ox, At//Oy.

The relational network of the objects in this problem can be represented as the figure below:



2.2 The kinds of facts

The kinds of facts in Relational Network of Geometric objects as follows:

(1) Fact of kind 1: information about object kind.

E.g.: [d,"LINE"] [O, "TRIANGLE"]

(2) Fact of kind 2: a determination of an object or an attribute of an object.

E.g.: SEGMENT[A,B] TRIANGLE[A,B,C].area

(3) Fact of kind 3: a determination of an object or an attribute of an object by a value or a constant expression.

E.g.: ANGEL[D,E,F] = $\frac{\pi}{3}$ TRIANGLE[A,B,C]. area = 12 cm^2

(4) Fact of kind 4: equality on objects or attributes of objects.

Eg.: SEGMENT[A,B] = SEGMENT[M,N]

(5) Fact of kind 5: a dependence of an object or an attribute of an object on other objects by a formula.

E.g.:
ANGEL[A,B,C]=ANGEL[A,B,D]+ANGEL[D,B,C]

(6) Fact of kind 6: a relation on objects or attributes of objects.

E.g.: ["BELONG",M,d]

["PERDENTICULAR",d1,d2]

(7) Fact of kind 7: equality between an object or an attribute and a function.

E.g.: M=INTESECTION(d1,d2)

N=MIDPOINT(A,B)).

3 THE ALGORITHM FOR THE PROBLEM OF AUTOMATIC DRAWING

3.1 Language specification for geometric problems

Each drawing object will be specified as a line which has structure as follow:

"**Draw** *OBJECT* **with** *FACTs;*" where,

– *OBJECT*: is a drawing object. It is specified as fact of kind 1 of ECOKB model.

– *FACTs*: is a set of facts to draw OBJECT. These facts are specified by kinds of fact of ECOKB model lacking of Ops.

Example 3.1: Draw angle xOy = 45º. Let A be in angle xOy. Draw line d1 through A and perpendicular to ray Ox at B. Draw line d2 through A and perpendicular to ray Oy at C.

Thus, this problem can be specified as follow:

"Draw [ANGLE[x,O,y], ANGLE] with [ANGLE[x,O,y].a=45];

Draw [A,POINT] with [IN,A,ANGLE[x,O,y]];

Draw [d1,LINE] with ["BELONG",A,d1],["PERPENDICULAR",d1,RAY[O,x]];

Draw [B,POINT] with [B=INTERSECTION(d1,RAY[O,x])];

Draw [d2,LINE] with ["BELONG",A,d2],["PERPENDICULAR",d2,RAY[O,y]];

Draw [C,POINT] with [C=INTERSECTION(d2,RAY[O,y])];"

3.2 Algorithm for automatic drawing

The algorithm for automatic drawing consists of 4 steps which divided into 2 stages:

Stage 1: Represent the relational network of objects from the hypothesis.

Step 1: Initiate and set initial states for some controlling variables.

Step 2: Check respectively the lines in the hypothesis; at each line, do:

If *OBJECT* has not been checked yet, do:

– Check the kind of *OBJECT*. For each kind of *OBJECT*, use Heuristic rules to generate new objects (if necessary) or new facts, and then update to *FACTs* set of the line.

– Put objects orderly into the vertices of the relational network.

Stage 2: Carry out drawing based on the relational network.

Step 3: Draw objects according to priority.

– *OBJECT* of primitive kind which has an empty *FACTs* set is generated freely.

– *OBJECT* of primitive kind which has a not empty *FACTs* set is drawn based on the facts of *FACTs* set.

– *OBJECT* of deductive kind is implicitly determined by previous objects.

All objects are determined by defining their coordinates or equations.

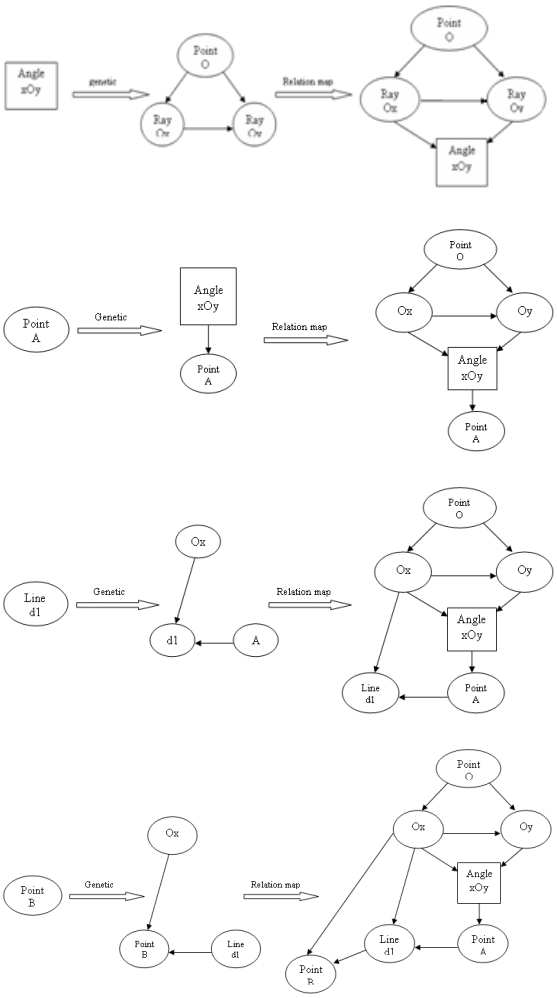
Step 4: Show the figure of the problem.

Example 3.1 (continue): with the above example, we have:

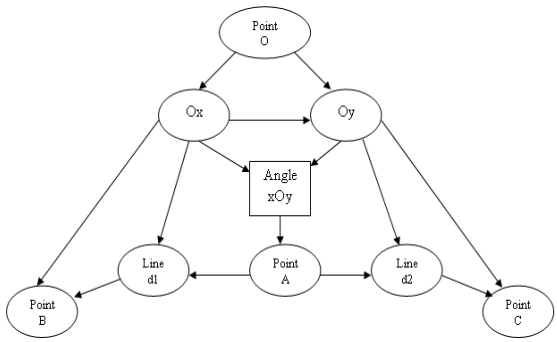
Stage 1:

Step 1: Initiate and set initial states for some controlling variables.

Step 2: Check the lines in the hypothesis; at each line, do:



Similarly to line d2 and point C, we have the relational network of the objects as follow:



Stage 2:

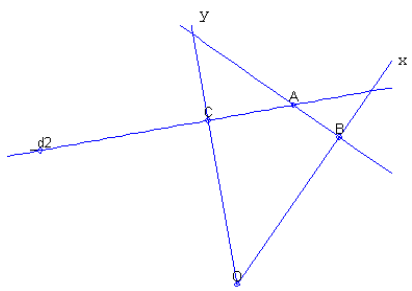
Step 3:

– Point O, ray Ox are drawn freely.

– Angle xOy is determined automatically after defining ray Oy.

– Other objects are determined based on the list of corresponding F facts.

Step 4: Show the figure of the problem.



4 APPLICATION

Based on [3], we used the relational network of geometric objects to design a module which can solve drawing problems in [6]. The module has a practical and important significance in solving geometry problems. In addition, this program is a systematic continuation of *the system which supports solving geometry problems in middle school* which was mentioned in [7]. As known, the system which supports solving geometry problems succeeded in solving reasoning problems which include proving or calculating. Nonetheless, a restriction of this system is that it does not have an illustrative figure for each problem. This makes the

reader unable to imagine the relations between geometric objects as well as the data from the problem. Therefore, the appearance of the program for automatic drawing is a perfect supplementation for the above system. The combining system assist readers easier to grasp the information and the relations between factors of a problem. In other words, such system will be a comprehensive tool with sufficient functions like human's (drawing and reasoning).

The drawing problems in [6] can be dealt with based on the following objects: POINT, LINE, RAY, CIRCLE (objects of primitive kind). The other objects such as SEGMENT, ANGLE, POLYGON, RAY (objects of deductive kind) will be determined automatically if the above objects (POINT, LINE, RAY, CIRCLE) is determined.

4.1 Drawing kinds of objects

Each drawing object will be specified as a line which has structure as follow:

4.1.1 POINT

- Point is defined freely.
- Point is in an angle (acute or obtuse) which is formed by two lines d1 and d2.
- Point is on a given line.
- Point is the central of a segment.
- Point is on a segment and forms with another point a given distance.
- Point forms with another point a segment which is parallel or perpendicular with another given segment.
- Point forms with another point a segment which has defined length.
- Point is the intersection of objects (LINE, SEGMENT or RAY).
- Point forms with previous points and facts to form a defined polygon (triangle, quadrangle, isoceles trapezoid, rectangle, lozenge, square, parallelogram).

4.1.2 LINE

- Arbitrary line.
- Line goes through a given point.
- Line forms with another line a defined angle.
- Line goes through a given point and parallel (or perpendicular) with another object.

4.1.3 RAY

Ray which is defined by two points is an object of deductive kind; thus it is determined automatically. Therefore, we just need check the ray which has only vertex is point.

- Ray is defined freely.
- Ray forms with another ray a defined angle.
- Ray forms with another ray an acute, right or obtuse angle.
- Ray is parallel or perpendicular with another object.

4.1.4 CIRCLE

- Circle in which its center and radius are defined.
- Circle goes through 3 vertices of a triangle

4.2 Application in solving drawing problems in middle school

The following examples show us clearly the effects which the program acquired when solving complicated drawing problems:

Example 4.1: Given two lines d1, d2 such that they form an angle of 60 degree. Let O be the intersection of d1 and d2. Let A be in the acute angle formed by d1 and d2. Draw segment AB such that B belongs to d1 and AB is perpendicular to d1. Draw segment BC such that C belongs to d2 and BC is perpendicular to d2.

Specification for the problem:

Hypothesis:=

"Draw [d1,LINE] with [ ];

Draw [d2,LINE] with [ANGLE[d1,d2]=60];

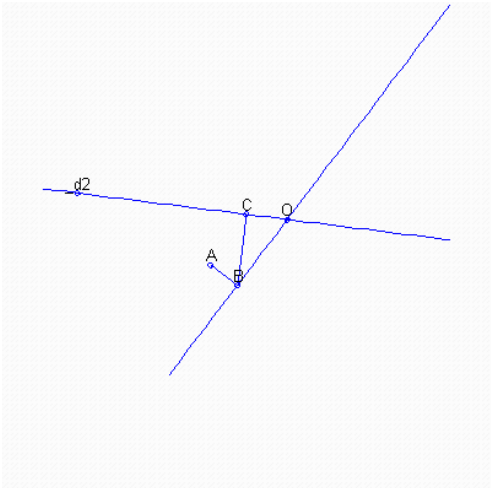
Draw [O,POINT] with [O=INTERSECTION(d1,d2)];

Draw [A,POINT] with [IN,A,ACUTEANGLE[d1,d2]];

Draw [SEGMENT[A,B],SEGMENT] with [PERPENDICULAR,SEGMENT[A,B],d1],[BELONG,B,d1];

Draw [SEGMENT[B,C],SEGMENT] with [PERPENDICULAR,SEGMENT[B,C],d2],[BELONG,C,d2];";

The result provided by the system:



Example 4.2: Given triangle ABC. Draw triangle A1B1C1 such that triangle A1B1C1 is congruent to triangle ABC by the ratio 2/3.

Specification for the problem:

Hypothesis:=

"Draw [TRIANGLE[A,B,C],TRIANGLE] with [ ] ;

Draw [TRIANGLE[A1,B1,C1],TRIANGLE] with [CONGRUENT,TRIANGLE[A1,B1,C1],TRIANGLE[A,B,C],2/3];";

The solution of the system:

– Draw triangle ABC freely.

Triangle A1B1C1 is congruent to triangle ABC by the ratio 2/3. So:

. A1B1 = 2/3*AB = 2/3 * a

. B1C1 = 2/3*BC = 2/3 * b

. A1C1 = 2/3*AC = 2/3 * c

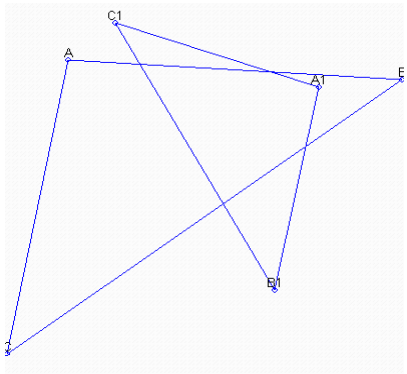
– Draw point A1 freely.

– Draw point B1 such that A1B1 = 2/3*a.

– Draw circle R1 which has center of A1 and radius of 2/3*c.

– Draw circle R2 which has center of B1 and radius of 2/3*b.

– Draw point C1 which is the intersection of R1 and R2.



Example 4.3: Draw quadrangle ABCD such that AB = 1.5, BC = 3, CD = 4, AD = 2 and angle ADC = 70 degree.

Specification for the problem:

Hypothesis:=

"Draw [QUADRANGLE[A,B,C,D],QUADRANGLE] with [SEGMENT[A,B]=1.5],[SEGMENT[B,C]=3], [SEGMENT[C,D]=4], [SEGMENT[A,D]=2],[ANGLE[A,D,C]=70];

";

The solution of the system:

– Draw angle xDy = 70.

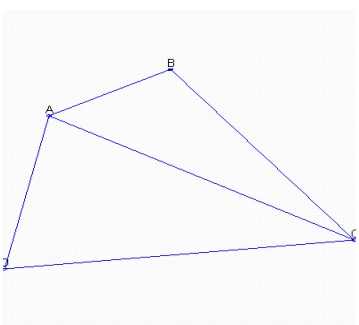
– Draw circle which has center of D, radius of 4 and intersects ray Dx at C.

– Draw circle which has center of D, radius of 2 and intersects ray Dy at A.

– Draw circle Q1 which has center of A and radius of 1.5.

– Draw circle Q2 which has center of C and radius of 3.

– Draw point B which is the intersection of Q1 and Q2.



5 CONCLUSION

The figure is an important component in geometric problem, so drawing figure automatically is very necessary for Intelligent Problems Solver system. Base on Computation Objects model and conceptual graph, the relational network between geometric objects is given, and algorithms to built it are too. By this relation network, the automatic drawing program geometric objects are designed succesfully. This program can draw four objects of primitive kind, include: POINT (9 cases), LINE (4 cases), RAY (5 cases), CIRCLE (2 cases). With these cases, the program can be draw almost the figure of problems in grade 6-7 and a part of problems in grade 8-9

The solution given by the automatic drawing program is natural, precise and has lively figure which helps the reader easier to grasp important information from the problem. In addition, this program and the system which supports solving geometry problems in middle school ([7, 12]) are combined to the system of solving geometry problems with the firgues and reasoning like human. Furthermore, this system has been estimated very well by the teachers and pupils used.

In the future, we will also continue researching work to draw the figure from the natural lanaguage. With this conversion, the communication with users and the system of solving geometry problems will become completely natural.

REFERENCES

1. Stuart Russell & Peter Norvig, *Artificial Intelligence – A modern approach* (third edition), Prentice Hall (2010).
2. John F. Sowa. *Knowledge Representation: Logical, Philosophical and Computational Foundations*, Brooks/Cole (2000).
3. Nhon Van Do (2012). *Intelligent Problem Solvers in Education: Design Method and Applications, Intelligent Systems*, Prof. Vladimir M. Koleshko (Ed.), ISBN: 978-953-51-0054-6, InTech.
4. Nhon Van Do, *Computational Networks for Knowledge Representation*, World Academy of Science, Engineering and Technology, Volume 56, August 2009, ISSN 2070 – 3724 (ICCSISE 2009), Singapore, 2009.
5. Nhon V. Do, Hien D. Nguyen, *A knowledge model about Relations and Application,* 4thInternational Conference on Data Mining and Intelligent Information Technology Applications (ICMIA 2012), pp. 701-704, Taiwan, October 2012.
6. Ministry of Education and Training, *Textbook and workbook of plane geometry of grades 6-7-8*, Publisher of Education (2006-2007).
7. Nhon Do, Hien Nguyen, *A Reasoning method on Knowledge Base of Computational Ojects and Designing a System for automatically solving plane geometry problems,* World Congress on Engineering and Computer Science 2011*,* (WCECS 2011)*,* San Francisco, USA, October 2011.
8. John F. Sowa, Arun K. Majumdar, *Analogical Reasoning*, Proceedings of the International Conference on Conceptual Structures in Dresden, Germany, in July 2003: A. Aldo, W. Lex, & B. Ganter, eds. (2003) *Conceptual Structures for Knowledge Creation and Communication*, LNAI 2746, Springer-Verlag, pp. 16-36.
9. W. Ertel, *Introduction to Artificial Intelligence*, ISSN: 1863-7310, Springer (2011).
10. George F. Luger & William A Stubblefield, *Artificial Intelligence*, Addison Wesley Longman, Inc (1998).
11. Frank van Harmelem & Vladimir & Bruce, *Handbook of Knowledge Representation*, Elsevier (2008).
12. Nhon V. Do, Hien D. Nguyen, Thanh T. Mai, *Designing an Intelligent Problems Solving System based on Knowledge about Sample Problems,* Proceeding of 5thAsian conference on Intelligent Information and Database Systems (ACIIDS 2013), Kuala Lumpur, Malaysia, March 2013, LNAI 7802, pp. 465-475, Springer.

THỂ LỆ VIẾT VÀ GỬI BÀI

1. Tạp chí Khoa học Trường Đại học Cần Thơ được Bộ Thông tin và Truyền thông cấp giấy phép số 200/GP.BTTTT ngày 29 tháng 5 năm 2013; mã số ISSN: 1859-2333 . Tạp chí thuộc danh mục các tạp chí được tính điểm công trình của một số Hội đồng chức danh Giáo sư nhà nước.

2. Tạp chí Khoa học Trường Đại học Cần Thơ được chia thành bốn (04) phần gồm (i) Khoa học Tự nhiên, Công nghệ và Môi trường; (ii) Nông nghiệp, Thủy sản và Công nghệ Sinh học; (iii) Khoa học Xã hội, Nhân văn và Giáo dục; và (iv) Khoa học Chính trị, Kinh tế và Pháp luật. Tạp chí công bố và giới thiệu các bài tổng quan (review paper), bài của các công trình nghiên cứu (research paper) và thông báo khoa học (short communication paper) của tác giả trong và ngoài nước. Nội dung công bố là các công trình chưa được xuất bản trên bất kỳ tạp chí nào khác. Tạp chí được xuất bản định kỳ 06 kỳ/năm.

3. Bài viết phải tuân thủ theo qui định của Tạp chí và được hướng dẫn chi tiết trên website: sj.ctu.edu.vn hoặc tapchikhoahoc.ctu.edu.vn

4. Tác giả gởi bài dạng tập tin (dạng tập tin MS. Word) cho Ban Biên tập (BBT) qua địa chỉ email: tapchidhct@ctu.edu.vn. Bài tổng quan không quá 30 trang, bài báo khoa học không quá 16 trang và bài thông báo khoa học (short communication) không quá 8 trang khổ A4, dòng đôi, kiểu chữ Times New Roman, kích thước 11. Bài viết không phải định dạng như qui định hiện hành. Khi gởi bài tác giả ghi rõ chọn phần nào của Tạp chí (dòng đầu tiên của trang thứ nhất).

5. Bài viết phải có phần tóm tắt bằng tiếng Việt và tiếng Anh khoảng 200 - 250 từ; có thêm tựa bài viết và các từ khóa bằng tiếng Việt và tiếng Anh.

6. Bài viết cần đánh dấu tác giả chịu trách nhiệm chính của bài viết với chi tiết về địa chỉ gởi thư, email, điện thoại. Tác giả đề xuất từ 01 đến 03 người phản biện cho bài viết (có thể trong hay ngoài Trường; có địa chỉ email và điện thoại kèm theo) là những người không tham gia vào bài viết (đồng tác giả, góp ý hay sửa bài viết) vào cuối bài viết. Ban Biên tập sẽ không gởi lại cho tác giả bài viết trong trường hợp bài không đạt yêu cầu của Tạp chí.

7. Ban Biên tập Tạp chí nhận bài thường xuyên, tổ chức thẩm định và xét duyệt theo quy định của Tạp chí. Các bài gởi đăng đạt yêu cầu sẽ được đăng trên số báo định kỳ phát hành gần nhất.

8. Địa chỉ gởi bài: Tòa soạn Tạp chí Khoa học Trường Đại học Cần Thơ
   Địa chỉ: Lầu 1, Khu Hiệu bộ, Khu II, Đường 3/2, Quận Ninh Kiều, Thành phố Cần Thơ
   Điện thoại: 0710 3 872 157; Fax: 0710 3 838 474
   Email: tapchidhct@ctu.edu.vn

9. Các vấn đề khác liên quan đến thể lệ gởi bài đăng xin vui lòng liên hệ:
   Thư ký Tòa soạn: ThS. Trần Thanh Điện
   Email: thanhdien@ctu.edu.vn hoặc tapchidhct@ctu.edu.vn