# TEXT SUMMARIZATION IN VARIOUS VIEWS AND ARBITRARY SCALE USING LATENT SEMANTIC INDEXING

Cao De Tran, Frederic Andres, Kinji Ono
National Institute of Informatics, Japan
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo, 101-8430, Japan
{tran,andres,ono}@nii.ac.jp

## ABSTRACT

**This paper proposes an approach to summarize a text document in arbitrary scale. We use Latent Semantic Indexing (LSI) [1] to deal with synonymy[1] and polysemy[2]. The text document is decomposed into a set of sentences and then summarized by a set of sentences that are high of the similarity to the view of user. A document may have various summaries for different user interests. The advantage of our method is that it is independent to the language used in the source text.**

**Our experimentation shows that the summary text can contains the sentences whose words are different to those used in the user view but their meaning are close to those used in the user view.**

## Keywords

Text summarization, synonymy, Latent Semantic, metadata extraction, information retrieval.

## 1. INTRODUCTION

Searching and retrieving information from large multimedia databases now is still an attracted issue in the literature. Despite a lot of effort has been done and many metadata extraction methods as well as searching and retrieving methods have been proposed, no method is good enough to find and retrieve a set of documents that contains just what the user needs. The users have to download the document, read it before deciding keep it or throw it away. This situation may become more trouble when they use a mall material like a cell phone or portable computer with small bandwidth and very limited capacity of memory and disk. Therefore we have a strong belief that a summary of multimedia content is as important as searching and retrieving them. In our research we investigate for the problem of summary multimedia document in view of user to offer a quick review on document content. User can describe his/her interest (hereafter, called *user view*) then the document will be summarized based on that view.

This paper proposes an approach to summarize a text document. The document is decomposed in a set of sentences. Each sentence is represented by a vector *term-by-sentence* [3]. The document is represented as a matrix whose column is a vector representing a sentence. The user describes his/her view by some keywords, so the user view can be considered as a query vector. To summarize the document, we find and retrieve *t* sentences that are closest to the user view from the document, where *t* is the number of sentences representing the desired scale. For example, if the document contains 100 sentences and the user want a summary in the scale of 10%, the summary will be made by 10 sentences. For this purpose we use a measure that calculates the similarity between the user view and the sentences in the document. The measure is used in this paper is cosine distance between two vectors [1,2].

To deal with the synonymy and polysemy, we use LSI, a technique based on the mathematical tool SVD (matrix' singular value decomposition) [1,2]. This technique is widely used to find and retrieve text documents and other kinds of media [5,6,7,8]. It is quite simple and powerful.

LSI was also applied to summary a text document by [3] or video content [4]. In [3], the textual document is summarized by a set of the sentences that cover the document content. Our work differs from the later because instead of doing an objective summary on global document content we would like to summarize according to user point of view. As a consequence, one document can have different summaries.

## 2. TEXT DOCUMENT MODEL

This part proposes a model for text document that serves to summarize the document. In our experiment, we implement two methods to summary a text document, both of them use the same text model that called *vector space model* [1,9]. One doesn't use LSI and the other uses LSI.

[1] Several words having the same or close meaning.

[2] Words having several meanings