

# A New LSI and TM based Cross-Language Information Retrieval System Providing Text Summaries

Liang Chen and Naoyuki Tokuda

Computer Science Department, Utsunomiya University, Utsunomiya, Japan, 321-8505

E-mail: lchen@alfn.mine.utsunomiya-u.ac.jp, tokuda@cc.utsunomiya-u.ac.jp

## Abstract

Traditionally information retrieval does not offer summaries, or summarize the documents on line after certain documents have been retrieved. It is not efficient for information retrieval in the internet, mainly because either inconvenient for the searchers for they need to read the retrieved texts or the system requires extremely too long time and computing facility. This paper describes a new concept IR system which offering the summaries just when the URL of the text is obtained. The Latent Semantic Indexing is suggested to be employed to index the summaries which might using different natural languages corresponding to the original texts, to offer the facility for cross-language retrieval, by avoiding the difficulties for translation of the query before searching. The translation memory technology is also applied to so that a rough translation could also be given on-line.

## Index Terms

Information Retrieval, Cross-language, summarization, Latent semantic index, key phrase

## I. INTRODUCTION

A so called search engine used in the internet is an information retrieval system in a traditional sense, responding to a keyword-based query typically with URLs as well as either title or closely related keywords associated with the web documents. As the number of documents in the net keeps increasing with such an explosive power, no one can afford time to read all of them to judge if they offer what we want. This motivates an advanced study on the new cross-language information retrieval system that can offer summaries of the web documents as well as the URL addresses in the native language we are acquainted with. An accurate and efficient summarization technology of documents constitutes the core of the new IR engines where the summaries extracted are used not only as a key element for information retrieval system, namely the template databases, but for online display to users together with the related URLs.

To make the summary data bases of the web pages, we use a fusion of the standard surface-level, entry-level and discourse-level approaches formally used in the literature, as well as fuzzy techniques and unsupervised learning schemes to improve the learning efficiency for adjusting the parameters, and the quality of summarization.

To help make the retrieval efficient for a longer queries, we make an extensive use of template structure we have developed for the ITS for language training. Based on the understanding that the closely matched sentences in the sense of the size of common sequence should be firstly used closely related phrases or words, i.e., the summaries matched to a query in the sense of longest common sequence (LCS) should be within the summaries close to the query in latent semantic space (figure). We firstly use LSI technique to limit the number of the summaries we need to check in the sense of LCS, before we use the algorithm checking LCS to find the real closely related summaries.