A study on Hierarchical Table of Indexes for Multi-documents

LE Thi Ngoc Tho (1010226)

School of Information Science,
Japan Advanced Institute of Science and Technology

August 09, 2012

Keywords: hierarchical summary, table of indexes, keyphrase extraction, clustering, unsupervised, graph based ranking.

Nowadays, when the information increase exponentially, catching up the new information is a time-consuming task for people, especially for busy ones. So, natural language processing is trying to support people in getting the news quickly by providing them a *summary* of text automatically, starting from summary of single document to multiple documents, or summary of news, meeting transcripts.

A summary of a document or a collection of documents is a condense representation of main ideas of the content. It is obvious that the summary of documents will help the readers gain the general ideas of documents. However, the representation of summary in form of text may cause inconvenience for the readers. Especially, the summary for a very long document or a collection of documents is still too long to read, and the non-native-speaker readers may not familiar with different writing styles. Even if the readers can get all ideas of documents, they have to figure out the structural organization of ideas by themselves.

In this thesis, we take into account the organization of main ideas as well when trying to get the summary of multiple documents. In order to do that, we generate a tree-based structure, called *hierarchical table of indexes*. A table of indexes in hierarchical structure helps the readers understanding

the content and the structure in semantics aspects. It also provides a navigation for the readers to quickly refer to interested information.

To create the hierarchical table of indexes automatically, we proposed an unsupervised framework to generate a hierarchical table of indexes. In which, unsupervised clustering algorithm is employed to create the hierarchical structure, and graph-based ranking method is applied to extract keyphrases and form the indexes. The experiment is applied for both English and Japanese in contribution to Legal Engineering. The preliminary result of summary is provided as the illustration for our approach. And searching information on the hierarchical summary is evaluated better than searching on original plain documents.