AN AUTOMATIC THAI TEXT SUMMARIZATION USING TOPIC SENSITIVE PAGERANK

Aekkasit Chongsuntornsri

A Thesis Submitted in Partial
Fulfillment of the Requirements for the Degree of
Master of Science (Computer Science)
School of Applied Statistics
National Institute of Development Administration
2007

AN AUTOMATIC THAI TEXT SUMMARIZATION USING TOPIC-SENSITIVE PAGERANK

Aekkasit Chongsuntornsri

School of Applied Statistics

The Examining Committee Approved This Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science (Computer Science).

Assistant Professor.....Committee Chairperson

(Ohm Sornil, Ph.D.)

Associate Professor.....Committee

(Pipat Hiranvanichakorn, D.E.)

Associate Professor .....Committee

(Surapong Auwatanamongkol, Ph.D.)

Associate Professor ..... Dean

(Surapong Auwatanamongkol, Ph.D.)

Date.....19 October 2007.....

ABSTRACT

Title of Thesis	An Automatic Thai Text Summarization Using Topic Sensitive PageRank
Author	Mr. Aekkasit Chongsuntornsri
Degree	Master of Science (Computer Science)
Year	2007

With the exponentially growing of information on the Internet and WWW, a large amount of the text collection expands. Automatic text summarization has become increasingly important to provide summary to help users retrieve the intended content rapidly. There are many applications of text summarization which include information extraction, knowledge acquisition, finding answers from specific questions, etc. The principle of automatic text summarization is that relevant information available in a document is abstracted into a shorter text by using a proper methodology.

This research proposes a technique for automatic Thai text summarization by using content-based and graph-based characteristics. It presents a logical view of texts by creating a relative document graph relying on content-based features. Topic Sensitive PageRank algorithm then assigns a score to each segment. A set of segments with high scores are extracted as a summary.

Experiments were conducted with Thai document data set and the DUC 2002 data set for English. There are two evaluation approaches to evaluate the proposed technique which is recall, precision and F1; and ROUGE measurement. The experimental results of the proposed technique show satisfactory results in the Thai and English documents summarization task.

ACKNOWLEDGEMENTS

I wish to thanks many people who give helpful, valuable contribution and suggestion that make this thesis possible. First of all, I would like to express my deepest appreciation to my major advisor, Assistant Professor Dr. Ohm Sornil, for valuable guidance, and for encouragement and support throughout the research. Moreover, I am deeply grateful to the committee members, Associate Professor Dr. Pipat Hiranvanichakorn and Associate Professor Dr. Surapong Auwatanamongkol, for their valuable advices and constructive comments.

Secondly, I also grateful to all faculty members of the School of Applied Statistics for their guidance and providing me knowledge and expertise. I also like to thank all officers of the School of Applied Statistics for their cooperation in providing related tasks for this thesis.

Thirdly, I would like to thank my companion, senior and junior collegian of the Department of Computer Science for their great suggestions and assistances.

Finally, I wish to express my gratitude to my parents and my sister for their supports, love, and understanding over the time.

Aekkasit Chongsuntornsri

October 2007