Universiti Teknologi MARA

Constructing Similarity Thesaurus for Military Documents

Suhaib Bin Said

Thesis Submitted In Partial Fulfillment of the Requirement for the
Bachelor of Computer Science (Hons.)
Faculty of Computer and Mathematical Sciences

July 2012

ACKNOWLEDGEMENT

Alhamdulillah, praise and thank to Allah because of His Almighty and His utmost blessings, I was able to finish this research within the time duration given. Firstly, my special thanks go to my supervisor, Muthukkaruppan Annamalai (Assoc. Prof. Dr.)for guiding me in completing this final year project and helping in me writing the thesis.

Special thanks also to other lecture who teach me during study in Uitm. Special appreciation also goes to my beloved parents who always support me and always encourage me. Last but not least, I would like to give my gratitude to my entire friend who helps me by giving some idea and some resource during the research.

ABSTRACT

This project is about constructing similarity thesaurus from collection of text documents. The collection that was use is collection of military an operation documents. Similarity thesaurus is group of word that share similar meaning. The major problem in information retrieval is word mismatch. This problem arose when the query word that use to search documents differ with word that been use in writing documents. Example the query word is money however the related documents use dollar instead money in the writing. So with similarity thesaurus help, the query will be expended to search document that related the dollar also. In this project 195 text documents were use. From 195 documents, 100 documents were using first to construct the similarity thesaurus. Then another 95 documents were added and the similarity thesaurus was compute again. The term to term relationship method was use in constructing the similarity thesaurus. The degree of confidence was introduced in this research to show the reliability of related term. Before the method applied word in the documents were tokenized. Then, removal punctual mark and stemming proceeds were applied to every token. Result show that size documents and collocation word in documents were variable that affect the similarity thesaurus. If more documents were use the more reliable and accurate the similarity thesaurus. The result also shows that the degree of confidence should be taken into account as well as the degree of similarity between two terms when grouping similar terms. Although the degree of similarity between two term are high however, if degree of confidence too low than the similarity between two term should not be taken.

Table of Contents

CHAPTER 1

INTRODUCTION

This chapter will present an overview of entire research dealing with similarity thesaurus. Presenting here is the research background, problems statement, objectives, scope and significance of this research.

1.1 Research Background

The purpose of information retrieval system is to assist user to find their required information item from large information's collection. This collection of information can be in different format such as documents, pictures and videos. Good information retrieval system must be able to provide accurate and relevant information item to user query. The main problem in information retrieval system is word mismatch(Abdulaziz, AbdulMalik, & Abdulrahman, 2009) . To retrieve a set of document, user should query to IR system with word that related to the document. For example he want document about money then he would required querying word related to money. Sometimes a problem will occur when there are some words that hold samilar meaning(Angel, Figuerola, & Berrocal, 2004). So, there will be the possibilities word that use by user for query differ from those used in writing documents(Angel, Figuerola, & Berrocal, 2004). For example the documents about dollar or currency will not be retrieving when the user write money in his query although these two words are related with each other. So the query should be reformulated by adding more terms or change it with suitable word to expend the query. Several approaches have been proposed by researchers to deal with this kind of problem. One of them is using thesaurus to expend the user query(Monica, 2002). A thesaurus is a collection of words in a given domain of knowledge and for each