

Query Expansion for Document Retrieval Using Wordnet

BY

MOHD HAFIZ BIN HUSSIN

BACHELOR OF COMPUTER SCIENCE (Hons)

**THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENT FOR BACHELOR OF COMPUTER
SCIENCE**

**FACULTY OF COMPUTER AND MATHEMATICAL
SCIENCES
UNIVERSITI TEKNOLOGI MARA**

OCT 2010

Acknowledgement

Firstly, I would like to express my gratitude towards Allah S.W.T that of his blessing that gives me capabilities, strength, and good health in order to finish up this project on time. In addition, this project successfully completed with my own effort and extra incentive. I would like to thank everybody who supported me with my project, especially my family for their moral support, love, advise and encouragement give me some confident through the progress of my research project.

Furthermore would like to special thank my thesis supervisor, Pn. Hayati Abd Rahman for encouraging, teaching, guiding, advising and supporting me to make sure that this project is successfully completed in time. Without the cooperation and her help, perhaps I have difficulties and I will not able to complete this project on time. Her patience in guiding me gives me confident to complete my project and I really appreciate it. I would also like to thank Dr. Noor Elaiza Binti Abd Khaled, our Final Year Project Coordinator for her guidance, support, encouragement, and criticism through the progress of my research project.

Last but not least, I would like to thank my friends, lecturers and everyone that is directly or indirectly involved in this project.

Thank you.

ABSTRACT

This paper describes the experimentation conducted to test the effectiveness of query expansion. Several experiments generating queries extracted from WordNet. Results show that lexical expansion is not able to improve retrieval performance. Nevertheless, the experiments allow us to conclude that query expansion can benefit searching process which allows structured queries.

KEYWORDS

Information retrieval, query expansion, sports, WordNet.

Table of Contents

| | |
|--|------------|
| DECLARATION | iii |
| ACKNOWLEDGEMENT | iv |
| ABSTRACT | v |
| TABLE OF CONTENTS | vi |
| LIST OF FIGURES | ix |
| LIST OF TABLES | x |
| LIST OF FORMULAS | x |
| | |
| 1. Chapter 1 - Introduction | 1 |
| 1.1 Background | 2 |
| 1.1.1 Relevant Document | 3 |
| 1.1.2 Query | 3 |
| 1.1.3 Query Expansion | 4 |
| 1.2 Problem Statement | 5 |
| 1.3 Objective | 6 |
| 1.4 Project Scope | 6 |
| 1.5 Significance | 7 |
| | |
| 2. Chapter 2 - Literature Review | 8 |
| 2.1 Introduction | 8 |
| 2.2 Query Expansion Experiments | 9 |
| 2.3 Query Expansion Technique | 12 |
| 2.3.1 Probabilistic Query Expansion | 13 |
| 2.3.2 Local Context Analysis | 14 |
| 2.4 Ideal Query Concept | 16 |
| 2.5 Wordnet | 18 |
| 2.5.1 Introduction to Wordnet | 18 |
| 2.5.2 Lexicon in Wordnet | 19 |
| 2.5.3 Semantic in Wordnet | 20 |
| 2.5.4 The Contents in Wordnet | 20 |
| 2.5.5 The Design of Wordnet | 21 |
| 2.5.6 Wordnet as Thesaurus | 22 |
| 2.5.7 Wordnet as a Dictionary | 23 |
| 2.5.8 Relations in Wordnet | 24 |
| 2.5.9 Sense of Disambiguation in Wordnet | 24 |
| | |
| 2.6 Information Retrieval with Wordnet | 25 |
| | |
| 2.7 Document Extraction for IR | 27 |

| | | |
|----------|---------------------------------------|----|
| 2.7.1 | Vector Space Model (VSM) | 28 |
| 2.7.2 | Term Weighting | 28 |
| 2.7.3 | Similarity Coefficients | 30 |
| 3. | Chapter 3 - Methodology | 31 |
| 3.1 | Introduction | 31 |
| 3.2 | Research Framework | 31 |
| 3.2.1 | Phase 1 : Analysis | 32 |
| 3.2.2 | Phase 2 : Design | 33 |
| 3.2.3 | Phase 3 : Data Collection | 34 |
| 3.2.4 | Phase 4 : Development | 34 |
| 3.2.5 | Phase 5 : Evaluation | 36 |
| 3.3 | Conceptual Framework | 37 |
| 3.3.1 | Information Retrieval Architecture | 37 |
| 3.3.2 | Research Framework | 38 |
| 3.3.2.1 | Query Keyword | 39 |
| 3.3.2.2 | Query Process | 39 |
| 3.3.2.3 | Query Representation | 40 |
| 3.2.3.4 | Corpus | 42 |
| 3.3.2.5 | Stemming Process | 43 |
| 3.3.2.6 | Stemming Algorithm | 44 |
| 3.3.2.7 | Removing Stop word Process | 44 |
| 3.3.2.8 | Problem in Removing Stop Word | 45 |
| 3.3.2.9 | Relevant Judgement | 45 |
| 3.3.2.10 | Document Representation | 46 |
| 3.3.2.11 | Matching Process | 47 |
| 3.3.2.12 | Listing the Document | 48 |
| 3.3.3 | Query Expansion | 49 |
| 3.3.4 | Redundancy in Wordnet | 50 |
| 3.4 | Recall and Precision | 51 |
| 3.5 | Software Requirement | 52 |
| 4. | Chapter 4 - Result and Finding | 53 |
| 4.1 | Introduction | 53 |
| 4.2 | Graphical Interface | 53 |
| 4.3 | Experimental Result | 56 |
| 4.3.1 | Query Number 1 | 57 |
| 4.3.2 | Query Number 2 | 58 |
| 4.3.3 | Query Number 3 | 59 |
| 4.3.4 | Query Number 4 | 60 |
| 4.3.5 | Query Number 5 | 61 |
| 4.3.6 | Query Number 6 | 62 |
| 4.3.7 | Query Number 7 | 63 |
| 4.3.8 | Query Number 8 | 64 |