Document Retrieval Using Semantic Approach

BY

MOHAMAD AFIF BIN ISMAIL

BACHELOR OF COMPUTER SCIENCE (Hons)

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

FACULTY OF COMPUTER AND MATHEMATICAL SCIENCES

UNIVERSITI TEKNOLOGI MARA

NOV 2010

Acknowledgement

Alhamdulillah, praise be to Allah the Almighty for giving me the strength and to finish this thesis. I would like to express my full gratitude to everyone who has helped me in doing this project.

I am heartily thankful to my supervisor, Hayati Abd Rahman, whose encouragement, guidance and support from the initial to the final level enabled me to successfully complete this project. Here I would also like to apologize for any wrongdoing that came from my part.

Secondly, thanks to my coordinators, Dr Noor Elaiza Binti Abd Khalid and Fakhrul Hazman Yusof for guiding me and CS2320 students in this subject.

I would like also to thank all the CS230 students for the collaboration that we had all over the year. Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the project.

Thank you very much

Abstract

As large amounts of digital information become more and more accessible, the ability to effectively find relevant information is increasingly important. Search engines have historically performed well at finding relevant information by relying primarily on lexical and word based measures. However, quite often these processes take place without respect to semantics, or word meanings. This is perhaps due to the fact that the idea of meaningful similarity is naturally qualitative, and thus difficult to integrate into quantitative processes. This project formally present a method for retrieving documents using semantic relations from WordNet database, which is designed to return document that have concept similarity with the user's query. This project show how semantic approach can be applied to document retrieval. Evaluation of the prototype being done used recall and precision method. The results conclude that while semantic is not well suited for precision of retrieval, the use of semantic approach can improve the number of document retrieved (recall).

Keywords: semantic, document retrieval, WordNet

Table of Contents

DECLARATION	**ii**
ACKNOWLEDGEMENT	**iii**
ABSTRACT	**iv**
TABLE OF CONTENTS	**v**
LIST OF FIGURES	**vi**
LIST OF ABBREVIATIONS	**vii**

1. **Chapter 1** – Introduction 1
   - 1.1 Background 1
   - 1.2 Problem Statement 2
   - 1.3 Objectives 3
   - 1.4 Project Scope 3
   - 1.5 Significance 4

2. **Chapter 2** – Literature Review 5
   - 2.1 Introduction 5
   - 2.2 Document Retrieval 6
   - 2.3 Document Representation 11
     - 2.3.1 Indexing Document 12
   - 2.4 Semantic Approach 13
   - 2.5 WordNet 14
     - 2.5.1 Introduction to WordNet 14
     - 2.5.2 Semantic in WordNet 15
   - 2.6 ConceptNet 18
   - 2.7 Evaluation 22
     - 2.7.1 Recall and Precision 22

3. **Chapter 3** – Methodology 23
   - 3.1 Introduction 23
   - 3.2 Development Framework 23
     - 3.2.1 Phase 1: Analysis 23
     - 3.2.2 Phase 2: Design 25
     - 3.2.3 Phase 3: Data Collection 26
     - 3.2.4 Phase 4: Development 26
     - 3.2.5 Phase 5: Evaluation 27
   - 3.3 Conceptual Framework 28
     - 3.3.1 Information Retrieval 29
     - 3.3.2 Search Engine 30
     - 3.3.3 Query Process 31

3.3.4 Query Representation 31
3.3.5 Removing Stopword 32
3.3.6 Stemming Process 34
3.3.7 Document Representation 35
3.3.7.1 Vector Space Model(VSM) 36
3.3.7.2 TF/IDF 37
3.3.8 Semantic Approach 38
3.3.9 Matching Function 41
3.4 Recall and Precision 43
3.5 Hardware and Software Requirement 44
3.5.1 Hardware Requirement 44
3.5.2 Software Requirement 44
3.6 Summary 44

4. **Chapter 4** – Result and Finding 45
4.1 Introduction 45
4.2 Graphical Interface 45
4.3 Evaluation Result 49
4.3.1 Query 1 49
4.3.2 Query 2 50
4.3.3 Query 3 51
4.3.4 Query 4 52
4.3.5 Query 5 53
4.3.6 Query 6 54
4.3.7 Query 7 55
4.3.8 Query 8 56
4.3.9 Query 9 57
4.3.10 Query 10 58
4.3.11 Graph 1 59
4.3.12 Graph 2 59
4.4 Summary 60

5. **Chapter 5** – Conclusion 61
5.1 Conclusion 61
5.2 Recommendation 62

References

Appendix A