Co-word Analysis Based Automatic Web Search

B.A.N.M. Bambarasinghe

09/10001

Faculty of Information Technology

University of Moratuwa

August 2011

Declaration

I declare that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a Degree or a Diploma in any University and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, to be made available for photocopying and for interlibrary loans, and for the title and summary to be made available to outside organization.

B.A.N.M. Bambarasinghe

Name of Student Signature of Student

Date

Supervised by

Prof. Asoka S. Karunananda

Name of Supervisor(s) Signature of Supervisor(s)

Date

Dedication

To all the scientists who cleared so many steps, so that we could clear one more

Acknowledgements

Our sincere gratitude is presented to professor A. S. Karunananda for his conceptual advices supported immensely to shape up the initial hypothesis, the feedback and various advices provided throughout the project which enhanced the quality of project and paper work. Prof. Priyan Dias is sincerely acknowledged for his clear teachings on philosophical background about knowledge and showing us the the light of connectionist approach for knowledge representation as the starting point of the project. Further more Dr. Shantha Jayalal is acknowledged for his invaluable knowledge sharing in the subject area of the project. All the members of Msc AI batch 2009 (university of Moratuwa) are acknowledged for strengthening the approach by raising questions and discussion points from different perspectives and support given throughout the conduct of the project.

Abstract

Automatic searching, knowledge acquisition and question answering are crying needs among the contemporary World Wide Web users. However conventional web is the major barriers for realizing above applications. This is because almost all important information in it is in natural languages and natural languages are very hard to be manipulated and understood by a computer. As a solution, more than a decade ago, semantic web was introduced, there was a lot of hope on machine understandability of the web. However the semantic web is still very far from realization due to the effort required for semantic tagging of available information.

In this project we try to build a solution for automatic searching in conventional web and similar information sources by mimicking the human natural language learning and knowledge representation process. Our approach is based on the hypothesis, which inspired by popular philosophy of science, that learning is matching known facts with new facts. In the context of conventional web, we employ statistical natural language processing technique co-word analysis for matching already available facts with new facts collected during automatic searching process.

As a proof of above hypothesis we have built a personalized automatic knowledge extraction application. That extracts knowledge from conventional web or similar information source regarding user queries and present synthesized documents related to the knowledge area of the query. Evaluation done by manual comparison of documents produced by the application and a document produced by a human user by web searching. Results showed automatic knowledge acquisition performs acceptable manner in most of the situations.

Contents

	Page
Chapter 1 Introduction	**1**
1.1 Introduction	1
1.2 Background and Motivation	1
1.3 Proposed Solution	3
1.4 Aim and Objectives	4
1.5 Summary	5
Chapter 2 Status of Automatic Knowledge Extraction	**6**
2.1 Introduction	6
2.1 Automatic knowledge extraction tools	6
2.2 Summary	12
Chapter 3 Knowledge Manipulation Techniques	**13**
3.1 Introduction	13
3.2 Co-word Analysis and Knowledge Extraction	13
3.3 Semantic Link Networks and Knowledge Representation	16
3.4 Rule Based Reasoning	21
3.5 Summary	23
Chapter 4 Approach for Automatic Knowledge Extraction	**24**
4.1 Introduction	24
4.2 Hypothesis	24
4.3 Personalized Knowledge Extraction System	25
4.4 Inputs and Outputs	25
4.5 Learning Process Using Co-word Analysis	27
4.5.1 Preprocessing of Source Documents	27
4.5.2 Co-word Analysis	27
4.5.3 Post Statistical Processing and Search Expansion	29
4.6 Synthesizing Documents Regarding User Queries	31
4.6.1. Identification of Sentences	31
4.6.2. Ranking of Sentences	32
4.6.3. Filtering Out Sentences	33
4.6.4. Reordering of Sentences	34
4.7 Useful Features	35
4.8 Summary	36

Chapter 5 Design of Automatic Knowledge Extractor **37**
5.1 Introduction 37
5.2 Design of Knowledge Extractor 38
5.3 Design of Knowledge Base 39
5.4 Design of Inference Engine 41
5.5 Design of User Interface 41
5.6 Summary 42
Chapter 6 Implementation of Automatic Knowledge Extractor **43**
6.1 Introduction 43
6.2 Software Development Platform and Core Technology 43
6.3 Implementation of Knowledge Extractor 44
6.3.1 Implementation of Knowledge Acquisition Sub System 44
6.3.2 Implementation of Knowledge Processing Sub System 45
6.4 Implementation of Knowledge Base 47
6.5 Implementation of Inference Engine 49
6.6 Implementation of User Interface 51
6.7 Summary 52
Chapter 7 How System Works **54**
7.1 Introduction 54
7.2 User Actions in Automatic Knowledge Extractor 54
7.3 Summary 54
Chapter 8 Evaluation **55**
8.1 Introduction 55
8.2 Evaluation Parameters 55
8.3 Experimental Setup 56
8.4 Analysis of Results 57
8.5 Summary 58
Chapter 9 Conclusion and Further Work **59**
9.1 Introduction 59
9.2 Discussion 59
9.3 Further Work 59
9.4 Summary 60
References **61**
Appendix A Knowledge Extractor Software Design Details **64**

List of Figures

	Page
Figure 1.1 - Proposed Automatic Knowledge Extraction System	4
Figure 4.1 - Inputs and Outputs of the System	27
Figure 4.2 - Knowledge Clusters	28
Figure 4.3 - Learning Process	31
Figure 4.4 - Example knowledge Cluster	32
Figure 4.5 - Synthesizing Process	35
Figure 5.1 - Top Level Architecture of Knowledge Extractor	37
Figure 5.2 - Knowledge Extractor Module	39
Figure 5.3 - Inference Engine Relationships	41
Figure 5.4 - User Interface Design	42
Figure 6.1 - Knowledge Base Database Schema	49
Figure 6.2 - Knowledge Extractor Search Interface	52
Figure 6.2 - Knowledge Extractor Explore Interface	53
Figure A.1 - Knowledge Extractor Class Diagram	63

List of Tables

	Page
Table 8.1 – Evaluation Results	57
Table 8.2 – Presentation Accuracy	58