

A CONCEPT SPACE APPROACH TO SEMANTIC EXCHANGE

by

Tobun Dorbin Ng

Copyright ©Tobun D. Ng 2000

A Dissertation Submitted to the

COMMITTEE on BUSINESS ADMINISTRATION

In Partial Fulfilment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY
WITH A MAJOR IN MANAGEMENT

In the Graduate College

THE UNIVERSITY OF ARIZONA

2 0 0 0

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: _____

ACKNOWLEDGMENTS

It has been an exceptional journey - one that I had never dreamed of. My research career was started not long after I became an MIS undergraduate student in 1989, when I attended Dr. Hsinchun Chen's first class at The University of Arizona. I am deeply indebted to Dr. Chen for his encouragement, advice, mentoring, and research support throughout my undergraduate, master's, and doctoral studies. I also truly appreciate his patience and tolerance during my numerous mishaps. This dissertation is part of the research carried out through his vision throughout last ten years.

I am fortunate to have the opportunity to work with a group of energetic people in Dr. Chen's AI Lab. I have enjoyed every moment that we have worked together including all those late night lab activities. All former and present AI Lab members have taught me many things about life. I appreciate all their friendships and their collective encouragement to finish this dissertation. I want especially to thank Chienting Lin, Thian-Huat Ong, Marshall Ramsey, Yohanes Santoso, Harry Li, Samuel Yim, and Chris Schuffels for sharing their technical wisdom, and to thank Andrea Houston, Joanne Martinez, Kris Tolle, and Rosie Hauck for sharing their research ideas.

Finally, it is impossible to have my research career without my parents' love and support, as well as my family and friends' encouragement. This dissertation is dedicated to them.

To all of you, thank you.

DEDICATION

This dissertation is a result of collective efforts from my parents, family, and friends. They often give me valuable advice, reminders, and wisdom about life, when I had a hard time understanding. However, they always give me room to explore and make mistakes.

This dissertation is dedicated to my parents, who have given me all their love and support and let me freely do whatever I want. Without them, there is no way I could possibly have accomplished this. Their understanding on the value of education is truly beyond my comprehension. I am just a lucky beneficiary. Along the way, my sister Karen and brother Gavin have shared their caring thoughts.

I have also been very lucky to have my uncle and aunt, Bob and Hannah Ng, and my cousins, Belinda and Gigi, throughout my stay in Arizona. They have helped me learn about life and live a good one. They have also given me comfort and advice whenever I needed them.

Throughout my ten plus years in Arizona and Dr. Hsinchun Chen's AI Lab, I have seen many people come and go; however, I am very grateful that many of them have taken me as their friend. They have taught me how to live, love, and feel; these, in turn, become the catalyst to my desire to finish this dissertation.

I truly thank all of them from the bottom of my heart.

We are loved beyond our capacity to comprehend. - Jewel Kilcher

TABLE OF CONTENTS

LIST OF FIGURES	10
LIST OF TABLES	11
ABSTRACT	12
1 INTRODUCTION	14
2 LITERATURE REVIEW	17
2.1 Static Nature of Knowledge in Library Collection	17
2.1.1 Characterizing Document Objects	18
2.1.1.1 Theory of Indexing	18
2.1.1.2 Manual Indexing	22
2.1.1.3 Indexing with Controlled Vocabulary and Thesauri	24
2.1.1.4 Free-text Indexing	25
2.1.2 Characterizing Global Knowledge in Document Collections	27
2.1.2.1 Classification Schemes and Categorization	27
2.1.2.2 Knowledge Discovery from Large Databases	30
2.1.2.3 Knowledge Bases, Inferencing, and Ontology	32
2.1.3 Revealing Knowledge in Neighborhood	34
2.1.3.1 Syntactic Mapping: Index List	34
2.1.3.2 Keyword Mapping: Controlled Vocabulary	35
2.1.3.3 Semantic Mapping: Man-made Thesauri	36
2.1.3.4 Co-occurred Mapping: Automatic Thesauri	37
2.2 Dynamic Nature of User Information Need	38
2.2.1 Expressing User Need	39
2.2.1.1 Information Need	39
2.2.1.2 Indeterminism	41
2.2.1.3 Opportunism	42
2.2.1.4 Vocabulary Problem	43
2.2.1.5 Recognition with Contextual Information	44
2.2.2 Perceiving Knowledge	45
2.2.2.1 Computing Relevance?	48
2.2.2.2 Structureless and Contextless: Document List	50
2.2.2.3 Structural but Contextless: Dynamic Clustering	51
2.2.2.4 Sturctural and Contextual: Path to the Knowledge	52

TABLE OF CONTENTS – *Continued*

3	RESEARCH QUESTIONS AND METHODOLOGIES	56
3.1	Research Questions	56
3.2	Methodologies	58
3.3	Dissertation Plan	60
3.3.1	A Framework for Semantic-enabled Information Technologies	60
3.3.2	Structure of Dissertation	67
4	CONCEPT SPACE CONSULTATION	69
4.1	Objectives	69
4.2	Research Questions and Methodology	71
4.3	Background and Issues	72
4.3.1	Spreading Activation in a Concept Network	72
4.3.2	Semantic Nets	73
4.3.3	Neural Nets	75
4.4	Concept Exploration in a Large Text-based Concept Network	78
4.4.1	Databases	79
4.4.2	Knowledge Discovery Methods	80
4.4.3	Knowledge Sources for Concept Network	82
4.4.4	Concept Exploration Methods	88
4.5	Two Algorithms for Spreading Activation	92
4.5.1	A Branch-and-bound Spreading Activation Algorithm: Semantic Net Based	92
4.5.2	A Hopfield Net Spreading Activation Algorithm: Neural Net Based	97
4.6	System Evaluation: A Benchmark Testing	102
4.7	User Evaluation: A Concept Exploration Experiment	103
4.7.1	Experimental Design	103
4.7.2	Experimental Results: Performance Analysis on Concept Retrieval	106
4.7.3	Experimental Results: Performance Analysis on Document Retrieval	114
4.8	Discussion and Conclusion	118

TABLE OF CONTENTS – *Continued*

5	LARGE-SCALE CONCEPT SPACE GENERATION	122
5.1	Objectives	122
5.2	Research Questions and Methodology	123
5.3	Background and Issues	124
5.3.1	Alleviating the Vocabulary Problem Using Concept Spaces .	124
5.3.2	Concept Association and Thesaurus Work	125
5.4	Concept Space Techniques	128
5.4.1	Document and object list collection	129
5.4.2	Object Filtering	130
5.4.3	Automatic Indexing	130
5.4.4	Co-occurrence Analysis	131
5.4.5	Associative Retrieval	134
5.5	Parallel Computing Approach	134
5.5.1	Parallel Computing for Information Retrieval	134
5.5.2	Parallel Computing for Knowledge Discovery	137
5.6	Parallel Computing: Concept Space Generation	139
5.6.1	Worm and Fly Concept Spaces Generation Using CM-5 and SGI Power Challenge	140
5.6.2	Computer Engineering Concept Space Generation Using SGI Power Challenge	144
5.6.3	CancerLit Concept Space Generation Using SGI Origin2000	146
5.6.4	Lessons Learned: Matching Concept Space Techniques with Parallel Computer Architecture	149
5.6.4.1	Data Parallelization	150
5.6.4.2	Algorithmic Parallelization	150
5.6.4.3	Synchronization: How Often?	152
5.7	System Evaluation: A Concept Association Experiment . .	152
5.7.1	Experimental Design	153
5.7.2	Experimental Results: Concept Recall and Concept Precision	154
5.8	Discussion and Conclusion	159

TABLE OF CONTENTS – *Continued*

6 CONCLUSIONS	162
6.1 Contributions	162
6.1.1 Dynamic Use of Knowledge Sources	162
6.1.2 Complementary to Man-made Thesauri	163
6.1.3 Context-specific Concept Spaces	164
6.1.4 Concept Space As Semantic Exchange Medium	166
6.2 Lessons Learned	168
6.3 Future Directions	170
Appendix A. BENCHMARK TESTING	173
A.1 Benchmark Testing Design	173
A.2 Benchmark Testing Results and Discussion	173
Appendix B. SAMPLE SESSIONS	177
Appendix C. CANCER SPACE: A WEB-BASED INFORMATION RETRIEVAL SYSTEM	182
C.1 Client User Interface	183
C.2 Server Implementation	186
Appendix D. FUNDING AND ACKNOWLEDGMENTS	189
D.1 Concept Space Consultation	189
D.2 Concept Space Generation	189
REFERENCES	191

LIST OF FIGURES

3.1	A framework for semantic-enabled information technologies	62
4.1	Frame-based representations for the knowledge sources	87
4.2	Connection weight assignment for the knowledge sources	91
4.3	Result of term recall on term retrieval	108
4.4	Result of term precision on term retrieval	109
4.5	Result of time spent on term retrieval	110
4.6	Result of contribution rate on term retrieval	111
4.7	Result of reviewing rate on term retrieval	112
4.8	Results of independent algorithms on document retrieval	115
4.9	Results of combined algorithms on document retrieval	116
5.1	CSQuest-suggested terms for “context analysis”	147
5.2	ANOVA analysis for recall and precision with graduate student subjects	158
5.3	ANOVA analysis for recall and precision with faculty subjects	158
6.1	Context-specific concept space consultation	165
C.1	Cancer Space: Results from concept space search	185
C.2	Cancer Space: Results from document retrieval	187

LIST OF TABLES

4.1	Terms/links in knowledge bases	86
5.1	A benchmark comparison summary on DEC Alpha, CM-5, and SGI Power Challenge	143
A.1	Results of benchmark testing	175

ABSTRACT

This dissertation work investigates the use of information technologies that clarify semantic meaning to help users elaborate their information needs by providing library-specific knowledge to the information seeking process. The research involved two interdependent semantic technologies: concept space consultation and library-specific, domain-specific, automatically generated concept spaces.

The concept space consultation phase used spreading activation algorithms - branch-and-bound and Hopfield net algorithms - to explore knowledge sources in specific domains. This research demonstrated the comparable effectiveness of exploration of a library database using a man-made classification scheme and thesaurus as opposed to an automatically generated concept space. The results showed that the use of spreading activation algorithms identified more relevant concepts than the use of the manual browsing method.

The concept space technique automatically identifies and extracts concept from a library collection while at the same time computing the strength of associations between concepts. This research demonstrated that the concept space technique was able to create human-recognizable concepts and their associations. In addition, the technique could be scaled to generate very large library-specific concept spaces for a very large underlying library collection.

Moreover, the interdependent use of both semantic technologies creates a semantic medium for users and library-specific knowledge sources to exchange content with context - context in user information need and that in corporeal knowledge.

CHAPTER 1

INTRODUCTION

This dissertation work investigated the use of information technologies that clarify semantic meaning to help users elaborate their information needs by specifying their library-specific knowledge during the information seeking process. That process starts with expression of a user information need, which makes it dynamic, cognitive, and user-oriented. However, existing information retrieval processes take only the end result of the cognitive aspect as query input and implicitly assume that the query input is the true (or almost true) representation of the user information need. Is this assumption always true?

The research in knowledge discovery, data mining, and machine learning has demonstrated the ability to generate knowledge of underlying data collections. In addition to automatic computational methods, extensive human effort has been invested in building domain-specific thesauri and classification schemes. However, all these semantic-bearing entities are targeted to represent knowledge coverage in particular domains and to capture semantic relationships between identified concepts. Can these knowledge sources adequately serve users' information needs?

When the information seeking process reaches the stage of retrieving information from search engines, users must perform two tasks - making a query and

evaluating a list of retrieved documents. If users are not satisfied with the results, they will repeat the same two tasks with different queries or they will give up their searches. This “blackbox” information retrieval approach has existed since the beginning of the information retrieval field in the 1960s and has since been popularized by the Internet. Nowadays, users are very familiar with what to expect from search engines - a possibly long ranked list of documents that they may never go through. In addition, in information retrieval practices, a common belief has been created that “relevant” documents appear only at the top 10 or 20 of ranked retrieved result.

The role of users in this “blackbox” approach focuses on evaluating retrieved documents for their relevancy. This approach downplays, if it does not ignore, the necessity of an information need having been fully expressed. This dissertation work adopted a user-centric and interactive approach to helping users elaborate their information needs with library-specific knowledge and simultaneously gain insight into a library’s offerings related to their information needs. Through semantic communication between users, users are enabled to express their information needs in the context of the library-specific knowledge of the target information source. Shared contextual information also gives users knowledge of the potential value of retrieved information. Under this approach, the information seeking process becomes a semantic journey from the initial expression of a user’s information need

to acquisition of knowledge of relevant information rather than a purely computational operation.

In order to explore semantic exchange between users and collections, this dissertation research focused on two areas: interactive consultation with knowledge sources and automatic generation of semantic-bearing knowledge sources from corresponding libraries. Concept space consultation demonstrates a semantic exchange between users and knowledge sources during query expression. Concept space generation shows how large-scale semantic-bearing knowledge sources can be automatically generated.