



UNIVERSITI PUTRA MALAYSIA

***KEYWORD QUERY PROCESSING INTERFACE MODEL OF ONTOLOGICAL
NATURAL LANGUAGE MANIPULATION***

SYED MUHAMMAD NOMAN HASANY

FK 2010 37

**KEYWORD QUERY PROCESSING INTERFACE
MODEL OF ONTOLOGICAL NATURAL LANGUAGE
MANIPULATION**

UPM By

SYED MUHAMMAD NOMAN HASANY

**DOCTOR OF PHILOSOPHY
UNIVERSITI PUTRA MALAYSIA**

2010

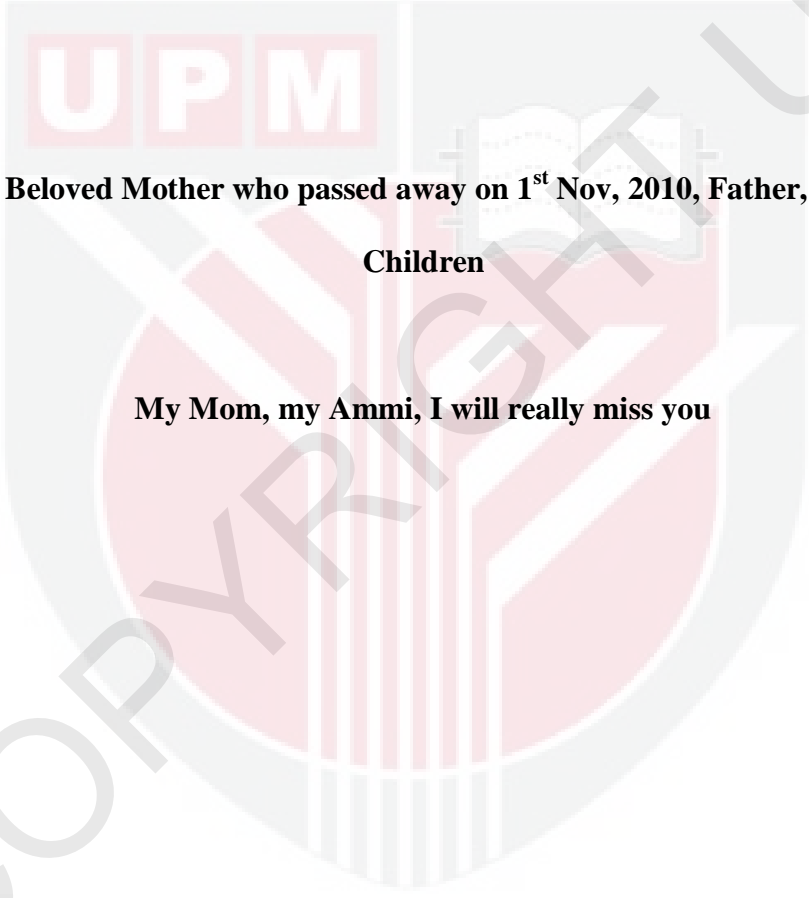
**KEYWORD QUERY PROCESSING INTERFACE MODEL OF
ONTOLOGICAL NATURAL LANGUAGE MANIPULATION**

By

SYED MUHAMMAD NOMAN HASANY

**Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in
Fulfilment of the Requirements for the Degree of Doctor of Philosophy**

October 2010



**To My Beloved Mother who passed away on 1st Nov, 2010, Father, Wife and
Children**

My Mom, my Ammi, I will really miss you

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

**KEYWORD QUERY PROCESSING INTERFACE MODEL OF
ONTOLOGICAL NATURAL LANGUAGE MANIPULATION**

By

SYED MUHAMMAD NOMAN HASANY

October 2010

Chairman: Associate Professor Adznan bin Jantan, PhD

Faculty: Engineering

Querying structured information through keyword queries provides an easy way to get to the information without knowing the structural details of the underlying data for formulating formal queries and without posing correct grammatical questions to the user interface. Besides the obvious advantages of keyword querying, it lacks expressiveness in contrast to syntactic questions. The problems faced by keyword queries lie in the fact that the processing capability is restricted to the posed keywords, additional connecting words and relations among keywords are ignored.

In semi-structured data like RDF, relations are formally defined as properties among concepts. This helps the keyword querying in finding connections among concepts from underlying data. But instead of this facility, the NLI results lack in precision and relevance. One major reason for this lacking is that more work is done in increasing

efficiency with respect to data storage, data indexing and reporting results using top-k strategies. Less work is performed in the direction of enhancing expressiveness, supporting lengthy queries and answering the queries with relevance oriented ranking.

We are concerned with enhancing the keyword query processing model in terms of handling expressive keyword queries and syntactic questions that incorporates quantifier restrictions and AND-OR semantics on RDF knowledge bases. The process of manipulating both type of natural language (NL) queries are supported by Ontologies. These NL queries are converted to target queries for result retrieval from RDF. The generated target queries are required to be ranked so that the results are reported in order to their relevance to the user query.

To handle large keyword queries, graph representation and processing is considered as a bottleneck. We preprocessed the RDF graph to be stored in distributed manner after the elimination of single chain productions in order to increase the efficiency in conversion process. We used the shortest path algorithms to be called on certain resources to explore connectivity to reduce complexity of search.

For the generality of target query representation and to incorporate quantifiers, subclasses and sub-class unions, we define an extended representation of the conjunctive query, termed as extended conjunctive query. But for the implementation of user query AND-OR semantics and semantic ranking, we define an efficient

representation, termed as compact Boolean query (CBQ). Empty result conditions reported by some approaches are also handled with the CBQ.

For the problem of conversion, techniques with fixed templates face scalability problems; while graph only techniques are processing intensive. We propose a variable template based conversion with inexpensive graph techniques to handle lengthy queries and exploring indirect connectivity among elements.

Considering the ranking problem, relevance ranking comprising of co-occurrence and Boolean semantics is proposed to help in understanding keyword queries and syntactic questions for precise answering.

Experimental results applied on LUBM, Mooney and self developed ontologies have shown that our technique can handle queries of 19 keywords within bearable time limits. The CBQ provides complete solution for empty results condition for correctly transformed queries. The coverage of queries is extended to understand queries originated from syntactic questions with improved precision. The improvement in values of MRR and TQP reflects the potential of our designed co-occurrence and AND-OR ranking strategies in placing the most relevant target queries at top positions.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk Ijazah Doktor Falsafah

MODEL ANTARAMUKA PEMROSESAN KATA KUNCI PENCARIAN YANG MEMANIPULASI BAHASA SEBENAR SECARA ONTOLOGI

Oleh

SYED MUHAMMAD NOMAN HASANY

Oktober 2010

Pengerusi : Profesor Madya Adznan bin Jantan, PhD

Fakulti : Kejuruteraan

Pencarian struktur maklumat melalui kata kunci pencarian adalah jalan mudah untuk mengetahui maklumat tanpa mengetahui struktur lengkap tentang data bagi menyusun pencarian yang formal, dan tanpa memerlukan persoalan tatabahasa yang betul pada antara muka pengguna. Di samping itu, kelebihan yang jelas pada kata kunci pencarian ini adalah kurangnya gaya ungkapan jika dibandingkan dengan susunan kata-kata dalam soalan. Masalah yang dihadapi oleh kata kunci pencarian ini bersandarkan pada fakta keupayaan pemprosesan yang terhad pada kata kunci yang tersedia, kata penghubung antara perkataan dan hubungan antara perkataan tidak dipedulikan.

Dalam separuh-struktur data seperti RDF, hubungan didefinisikan secara formal sebagai ciri-ciri/sifat antara konsep. Ini membantu pencarian menggunakan kata kunci dalam mencari kaitan antara konsep data. Tetapi, walaupun dengan kemudahan ini, keputusan NLI adalah kurang dari segi ketelitian dan perkaitan. Salah satu factor utama

kekurangan ini adalah kerja yang lebih diperlukan untuk meningkatkan kecekapan dengan penyimpanan data, pengindeksan data, dan keputusan laporan menggunakan strategi *top-k*. Hanya sedikit kerja yang telah dijalankan dalam pencarian memperluaskan gaya ungkapan, pendukung pencarian yang panjang dan menjawab pencarian dengan relevan berorientasikan kedudukan.

Kami menitik beratkan dengan memperluaskan model pemrosesan kata kunci pencarian dari sudut gaya ungkapan kata kunci pencarian dan soalan-soalan sintetik yang merangkumi sekatan penjangka dan semantik DAN-ATAU dalam pangkalan pengetahuan RDF. Proses memanipulasi kedua-dua jenis pencarian menggunakan bahasa sebenar (NL) adalah didukung oleh Ontologi. Pencarian NL ini ditukar menjadi pencarian sasaran untuk mendapatkan keputusan dari RDF. Sasaran pencarian yang dibina diperlukan untuk digolongkan/disusun supaya hasil yang dilaporkan mempunyai perkaitan dengan pencarian pengguna.

Untuk menguruskan kata kunci pencarian yang besar, perwakilan graf dan pemrosesan diambil kira sebagai perkara utama. Kami melakukan pra-proses graf RDF untuk diletakkan dalam pembahagian yang betul selepas pengeluaran rantai tunggal dihapuskan bagi meningkatkan kecekapan proses penukaran. Kami menggunakan algoritma yang ringkas untuk dipanggil oleh sumber tertentu bagi menjelajah perkaitan antara perkataan untuk mengurangkan pencarian yang kompleks.

Untuk sasaran pencarian yang umum dan menggabungkan penjangka, sub-kelas dan kesatuan sub-kelas, kami menetapkan perluasan perwakilan bagi pencarian kata penghubung, disebut sebagai perluasan pencarian kata hubung. Tetapi untuk pelaksanaan pencarian pengguna menggunakan semantik DAN-ATAU dan semantik kedudukan, kami menetapkan perwakilan yang cekap, disebut sebagai pencarian Boolean padu (CBQ). Keadaan keputusan 'tiada hasil' yang dilaporkan oleh beberapa pendekatan juga telah ditangani menggunakan CBQ.

Bagi permasalahan penukaran, teknik dengan templat tetap berdepan dengan masalah pengukuran, manakala teknik dengan graf sahaja memproses secara intensif. Kami mencadangkan penukaran pangkalan pembolehubah templat dengan teknik graf mudah untuk menangani pencarian yang panjang dan menjelajah kaitan yang tidak langsung antara elemen.

Mengambil kira masalah kedudukan, kesesuaian kedudukan yang mengandungi kemunculan awal dan sematik Boolean dicadangkan untuk membantu dalam memahami kata kunci pencarian dan soalan-soalan sintetik untuk jawapan yang tepat.

Keputusan eksperimen ke atas LUBM, Mooney dan ontologi ciptaan sendiri telah menunjukkan yang teknik kami boleh menangani pencarian dengan menggunakan 19 kata kunci dengan penggunaan batas masa yang baik. CBQ menyediakan penyelesaian lengkap untuk keadaan keputusan 'tiada hasil' bagi mengubah pencarian dengan betul. Liputan pencarian juga dapat diperluaskan untuk memahami pencarian yang asalnya

daripada susunan soalan dengan meningkatkan ketelitian. Pembaikan pada nilai MRR dan TQP adalah kesan daripada potensi rekaan kejadian-berulang kami dan strategi kedudukan DAN-OR dalam menyusun sasaran pencarian yang paling relevan pada kedudukan paling atas.



ACKNOWLEDGEMENTS

I thank ALLAH Almighty for all things throughout my voyage of knowledge exploration.

I would like to express my sincere gratitude to my supervisor Associate Professor Dr. Adznan Bin Jantan and also my supervisory committee members Associate Professor Mohd. Hasan Selamat and Dr. Iqbal Saripan for their guidance and advice throughout this work in making this a success.

My special thanks to Associate Professor Mohd. Hasan Selamat for his technical and financial help in the development of the prototypes and for providing his high-speed server for testing and evaluation works.

My deepest appreciation to my mother who passed away on 1st November, 2010, my family, my father for their utmost support and encouragement without which all these would not be possible. Special thanks to my wife who suffered the most during this period.

For the others who have directly or indirectly helped me in the completion of my work, I thank you all.

APPROVAL

I certify that an Examination Committee met on 14th October, 2010 to conduct the final examination of Syed Muhammad Noman Hasany on his Doctor of Philosophy thesis entitled "**Keyword Query Processing Interface Model Of Ontological Natural Language Manipulation**" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee were as follows:

Associate Professor
Department of Computer and Communication Systems Engineering
Faculty of Engineering
Universiti Putra Malaysia
(Chairman)

Associate Professor
Department of Computer and Communication Systems Engineering
Faculty of Engineering
Universiti Putra Malaysia
(Member)

Associate Professor
Department of Computer and Communication Systems Engineering
Faculty of Engineering Universiti Putra Malaysia
(Member)

(External Examiner)
Associate Professor

BUJANG KIM HUAT, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 26th November, 2010

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Adznan bin Jantan, PhD

Associate Professor
Faculty of Engineering
Universiti Putra Malaysia
(Chairman)

Mohd. Hasan Selamat, PhD

Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

M. Iqbal Saripan, PhD

Lecturer
Faculty of Engineering
Universiti Putra Malaysia
(Member)

HASANAH MOHD, GHAZALI, PhD

Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

TABLE OF CONTENTS

	Page
DEDICATION	iii
ABSTRACT	iv
ABSTRAK	vii
ACKNOWLEDGEMENTS	xi
APPROVAL	xii
LIST OF TABLES	xviii
LIST OF FIGURES	xix
LIST OF ABBREVIATIONS	xxiii
CHAPTER	
1 INTRODUCTION	1
1.1 Background	1
1.1.1 Semantic Web	3
1.1.2 Resource Description Framework	5
1.1.3 RDF, RDFS and XML	7
1.1.4 Ontology and RDF Knowledge Base	8
1.1.5 RDF as a Semantic Web Document	8
1.1.6 Natural Language Interface	9
1.1.7 Keyword Query Processing Interface	10
1.2 Motivation of Study	11
1.3 Problem Statement	12
1.4 Research Objectives	16
1.5 Assumption	17
1.6 Scope of Study	18
1.7 Thesis Organization	18
2 LITERATURE REVIEW	21
2.1 Introduction	21
2.2 NLI to Structured and Semi-structured Data	21
2.2.1 NLI to Databases	22
2.2.2 NLI to XML	23
2.2.3 NLI to Ontological KBs	23
2.3 Sub-problems for NLI to Ontologies	25
2.3.1 Input Format Problem	25
2.3.2 Query Conversion Problem	30
2.3.3 Target Query	37
2.3.4 Formal Query Expressivity Problem	39

2.3.5	Query Length	40
2.3.6	The Granularity Problem	41
2.3.7	Ranking	41
2.3.8	Evaluation	43
2.4	Summary	44
3	METHODOLOGY	49
3.1	Introduction	49
3.2	Research Methodology	49
3.2.1	Research steps	49
3.2.2	The prototype: QuriOnto	53
3.2.3	The benchmark system: SPARK-base	58
3.2.4	The Data set	59
3.2.5	Evaluation parameters	60
3.3	Four Stage NLI Model	61
3.3.1	Preprocessing Stage	62
3.3.2	Keyword Query Stage	62
3.3.3	Conversion Stage	64
3.3.4	Ranking Stage	66
4	QURIONTO: THE ENHANCED KEYWORD QUERY PROCESSING INTERFACE FOR RDF	67
4.1	Introduction	67
4.2	QuriOnto	68
4.3	RDF Graph Preprocessing	70
4.3.1	Single Chains Suppressed Graph	70
4.4	Query Level Processing	76
4.4.1	Quantifier Restrictions Processing	76
4.4.2	Shallow Linguistic Processing	79
4.4.3	Mapping Query terms to Ontological resources	80
4.5	Conversion Level Processing	81
4.5.1	Query Division for Processing	83
4.5.2	Conversion Structures	84
4.5.3	Graph Connectivity Algorithm	89
4.5.4	Description of Conversion Algorithms	92
4.5.5	An Illustrative Example of Algorithm	103
4.6	QuriOnto Intermediate Target Queries	111
4.6.1	Extended Conjunctive Query	111
4.6.2	Compact Boolean Query	112
4.6.3	ECQ to CBQ Transformation	115
4.6.4	CBQ to SPARQL	127
4.7	Ranking	128
4.7.1	Mapping/Matching Proximity Scoring	131
4.7.2	Path Length Scoring	131
4.7.3	Co-occurrence Ranking	132
4.7.4	AND-OR Ranking	136

4.8	SPARK-base Implementation	137
4.8.1	Resources Identification	137
4.8.2	Query Conversion	137
4.8.3	Ranking	138
4.9	Evaluation Parameters	138
5	QURIONTO EVALUATION	141
5.1	Introduction	141
5.2	Prototype Implementation	143
5.3	Dataset	144
5.4	Precision	146
5.4.1	Query Size Scaling with Resources	147
5.5	Ranking Effectiveness	149
5.5.1	Mean Reciprocal Rank	150
5.5.2	Target Query Position	152
5.6	Efficiency	156
5.7	Coverage	158
5.7.1	AND-OR Semantics	158
5.7.2	QuriOnto Evaluation for Quantifier Restrictions	159
5.7.3	Empty Results Condition	160
5.7.4	Answer Granularity	160
6	CONCLUSIONS AND FUTURE WORK	162
6.1	Introduction	162
6.2	Limitations and Future Work	164
6.3	Contributions to Knowledge	164
6.4	Conclusions	165
	REFERENCES	167
	APPENDIX A	178
	APPENDIX B	180
	APPENDIX C	181
	APPENDIX D	182
	BIODATA OF STUDENT	184