UNIVERSITI PUTRA MALAYSIA

EFFECTIVE QUERY STRUCTURING WITH RANKING USING NAMED ENTITY CATEGORIES FOR XML RETRIEVAL

ABUBAKAR ROKO

FSKTM 2016 18

EFFECTIVE QUERY STRUCTURING WITH RANKING USING NAMED ENTITY CATEGORIES FOR XML RETRIEVAL

By

ABUBAKAR ROKO

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

September 2016

DEDICATION

To my mother for without her support this thesis would not have been possible

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirement for the Degree of Doctor of Philosophy

EFFECTIVE QUERY STRUCTURING WITH RANKING USING NAMED ENTITY CATEGORIES FOR XML RETRIEVAL

By

ABUBAKAR ROKO

September 2016

Chairman : Associate Professor Shyamala Doraisamy, PhD
Faculty : Computer Science and Information Technology

A large number of documents are now represented and stored using an XML document structure on the web. Thus, there is a need for effective and user-friendly search systems for XML document search. Query languages are largely used to compose structured queries by users to extract data from XML documents. However, using query languages to express queries prove to be difficult for most users since this requires learning a query language and knowledge of the underlying data schema. On the other hand, the success of Web search engines has made many users to be familiar with keyword search and therefore prefer to use a keyword search query interface to search XML data. Keyword queries are inherently ambiguous and it is difficult for users to clearly state their intentions, which causes keyword search systems to inevitably return irrelevant results, making search engines less effective. Therefore, to improve the effectiveness of search engines, keyword search systems are highly needed.

Query structuring system is one of the keyword search systems recently used for effective retrieval of XML documents. The systems focus on user query representation, user search intention identification and ranking algorithms to improve keyword search. However, firstly, existing systems return wrong query representation because of their inability to put keyword query ambiguity problems into consideration during query pre-processing. For example, none of the systems consider the following ambiguities: (i) a query term can appear as the text values of different XML nodes and having different semantics (ii) a query term can appear as both a tag name and as part of text content of some node. Secondly, the systems return wrong user search intention. Specifically, the systems return irrelevant predicates as well as non-informative entity nodes. Thirdly, the systems fail to generate and select best structured query that match a user input keyword query. Finally, the systems' ranking functions ignore to consider the semantics of XML tags into account which leads to irrelevant results. These problems are addressed as follows:

Firstly, an enrichment method has been proposed to investigate whether enriching document content with semantic tags improves the performance of keyword queries. The method employs Semantic Tags Extraction (STSE) algorithm to extract semantic tags of an element and Element Enrichment (EERM) algorithm to enrich the elements.

Secondly, a XML Keyword Query Structuring System (XKQSS) has been developed to relegate the task of generating structured queries from a user to itself while retaining the simple keyword search query interface that allows users to submit a schema independent keyword query. The XKQSS uses a Semantic Aware Index scheme (SAIS) to record the proportion of Named Entity Categories (NECs) and an Entity based Query Segmentation (EBQS) method to interpret the user query as a list of keywords and named entities (resolves ambiguity). Furthermore, it employs Predicates Identification Algorithm (PIA) and Entity Identification Algorithm (EIA) to identify user search intention. Finally, the system utilizes a query formulation algorithm (QRYF) to select the structured queries that best interpret user query.

Thirdly, a modification to XKQSS called Ranking Aware XML Keyword Query Structuring System (RAXKQSS) has been developed to effectively return a ranked list of elements as answer to a user query. The RAXKQSS, first, introduces an improve SAIS (ISAIS) to record the Named Entity Category (NEC) of each indexed term, in addition to the usual information such as term frequencies, term position, as well as element that contains the term in the inverted index. Then, the system uses a ranking function *rk*_BM25TOPF to assign relevance scores to XML fragments with respect to a query and an N-gram based Query Segmentation (NBQS) method to interpret the user query as a list of N-grams (resolves ambiguity). Next, it introduces an Improved PIA (IPIA) and a Compute Return Node Algorithm (CRNA) to return relevant predicates and return node, respectively. Finally, the system employs a query formulation via node algorithm (QRYFv) algorithm to improve the selection of structured queries that best match user query

Experiments have been conducted to evaluate the performance of the proposed enrichment method, XKQSS and RAXKQSS. The experimental results have shown that the enrichment method has an insignificant improvement compared with the baseline in terms of Mean Average Precision (MAP). The results also demonstrated that the propose XKQSS outperforms XReal and StruX in terms of precision. Moreover, the results also illustrated that the proposed RAXKQSS achieved higher precision when compared with the StruX, the SLCA.

These results have shown that the enrichment method is ineffective in improving retrieval performance while the proposed systems XKQSS and RAXKQSS have proved effective compared to the StruX and the SLCA in terms of retrieval performance.

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

PERTANYAAN PENSTRUKTURAN BERKESAN DENGAN KEDUDUKAN MENGGUNAKAN KATEGORI ENTITI DINAMAKAN UNTUK DAPATAN SEMULA XML

Oleh

ABUBAKAR ROKO

September 2016

Pengerusi : Profesor Madya Shyamala Doraisamy, PhD
Fakulti : Sains Komputer dan Teknolologi Maklumat

Sebilangan besar dokumen kini diwakili dan disimpan dengan menggunakan struktur dokumen XML di laman web. Oleh itu, terdapat keperluan sistem carian yang berkesan dan mesra pengguna bagi carian dokumen XML. Bahasa pertanyaan sebahagian besarnya digunakan untuk mengarang pertanyaan berstruktur oleh pengguna untuk mengekstrak data dari dokumen XML. Walau bagaimanapun, menggunakan bahasa pertanyaan untuk menyatakan pertanyaan terbukti sukar bagi kebanyakan pengguna kerana ini memerlukan mereka untuk belajar bahasa pertanyaan dan pengetahuan skema data asas. Sebaliknya, kejayaan enjin gelintar laman web telah menyebabkan ramai pengguna untuk membiasakan diri dengan carian kata kunci dan seterusnya lebih cenderung untuk menggunakan carian pertanyaan kata kunci antara muka untuk carian data XML. Pertanyaan kata kunci sememangnya tidak jelas dan ia adalah sukar bagi pengguna untuk menyatakan dengan jelas niat mereka menyebabkan enjin gelintar kata kunci untuk menjana hasil yang tidak relevan, menjadikan enjin gelintar kurang berkesan. Oleh itu, untuk meningkatkan keberkesanan enjin gelintar, sistem carian kata kunci adalah sangat diperlukan.

Sistem penstrukturan pertanyaan adalah salah satu sistem carian kata kunci baru-baru ini digunakan bagi dapatan semula dokumen XML yang berkesan. Sistem yang sedia ada memerlukan pengguna untuk memberikan petunjuk berstruktur dalam input pertanyaan kata kunci mereka walaupun antara muka sistem ini adalah kata kunci asas. Limitasi lain sistem ini juga termasuk, pertama ketidakupayaan untuk mempertimbangkan meletakkan masalah kekaburan pertanyaan kata kunci semasa pra-pemprosesan pertanyaan. Sebagai contoh, tiada sistem mempertimbangkan kekaburan berikut: (i) istilah pertanyaan boleh muncul sebagai nilai teks bagi nod XML berbeza dan mempunyai semantik yang berbeza (ii) istilah pertanyaan boleh muncul sebagai tag nama dan sebahagian daripada kandungan teks sesetengah nod. Kedua, sistem memulangkan hasrat carian pengguna yang berlainan. Ketiga, sistem gagal untuk menjana dan memilih pertanyaan berstruktur terbaik yang sepadan dengan pertanyaan kata kunci yang diberikan. Akhir, sekali fungsi kedudukan sistem

mengabaikan pertimbangan tag semantik XML yang membawa kepada hasil yang tidak relevan. Masalah-masalah ini ditangani seperti berikut:

Pertama, kaedah pengayaan telah dicadangkan untuk mengkaji sama ada memperkaya kandungan dokumen dengan tag semantik meningkatkan prestasi pertanyaan kata kunci. Kaedah ini menggunakan algoritma Pengekstrakan Tag Semantik (STSE) untuk mengekstrak tag semantik elemen dan algoritma Elemen Pengayaan (EERM) untuk memperkayakan elemen-elemen.

Kedua, Sistem Pertanyaan Kata Kunci Berstruktur (XKQSS) telah dibangunkan untuk membuang tugas menjana pertanyaan berstruktur daripada pengguna kepada dirinya dalam masa yang sama mengekalkan carian pertanyaan kata kunci antara muka mudah yang membolehkan pengguna untuk mengemukakan skema kata kunci pertanyaan bebas. XKQSS menggunakan Skim Indeks Mengetahui Semantik (SAIS) untuk merakam perkadaran Kategori Entiti Dinamakan (NECs) dan kaedah Entiti berdasarkan Segmentasi Pertanyaan (EBQS) untuk mentafsir pertanyaan pengguna sebagai senarai kata kunci dan entiti yang dinamakan (mengatasi kekaburan). Tambahan pula, ia menggunakan Algoritma Pengenalan Predikat (PIA) dan Algoritma Pengenalan Entiti (EIA) untuk mengenal pasti hasrat carian pengguna. Akhirnya, sistem ini menggunakan Algoritma Formulasi Pertanyaan (QRYF) untuk memilih pertanyaan berstruktur yang terbaik mentafsir pertanyaan pengguna.

Ketiga, satu pengubahsuaian kepada XKQSS dipanggil Sistem Kedudukan Sedar Pertanyan Kata Kunci Berstruktur (RAXKQSS) telah dibangunkan bagi mendapatkan semula secara berkesan satu senarai elemen tersusun sebagai jawapan kepada pertanyaan pengguna. The RAXKQSS, pertamanya memperkenalkan SAIS diperbaik (ISAIS) untuk merakam Kategori Entiti Dinamakan (NEC) bagi setiap istilah diindeks, sebagai tambahan kepada maklumat biasa seperti kekerapan istilah, kedudukan istilah, dan juga elemen yang mengandungi istilah yang ada pada indeks songsang. Kemudian, sistem menggunakan fungsi kedudukan *rk*_BM25TOPF untuk menetapkan skor kaitan fragmen XML berkenaan dengan pertanyaan dan N-gram berdasarkan kaedah Segmentasi Pertanyaan (NBQS) untuk mentafsir pertanyaan pengguna sebagai senarai n-gram (mengatasi kekaburan). Seterusnya, ia memperkenalkan PIA diperbaik (IPIA) dan Algoritma Pengiraan Pulangan Nod (CRNA) untuk mengembalikan predikat relevan dan nod kembali, masing-masing. Akhirnya, sistem ini menggunakan algoritma Formulasi Pertanyaan melalui Algoritma Nod (QRYFv) untuk meningkatkan pemilihan pertanyaan berstruktur yang terbaik menandingi pertanyaan pengguna.

Eksperimen telah dijalankan untuk menilai prestasi kaedah pengayaan yang dicadangkan, XKQSS dan RAXKQSS. Hasil eksperimen telah menunjukkan bahawa kaedah pengayaan mempunyai peningkatan yang tidak ketara berbanding garis asas dari segi Min Purata Persis (MAP). Hasil juga menunjukkan bahawa XKQSS yang dicadangkan melebihi performa XReal dan StruX dari segi ketepatan. Selain itu, hasil juga menggambarkan bahawa RAXKQSS yang dicadangkan mencapai ketepatan yang lebih tinggi jika dibandingkan dengan StruX dan SLCA.

Hasil ini menunjukkan bahawa kaedah pengayaan adalah tidak berkesan dalam meningkatkan prestasi dapatan semula manakala sistem XKQSS dan RAXKQSS yang

dicadangkan telah terbukti berkesan berbanding dengan StruX dan SLCA dari segi prestasi dapatan semula.

ACKNOWLEGDEMENTS

First of all, I wish to praise Allah *Subhanahu Wa Taala* for endowing me with courage, patience, and guidance to complete this study.

Secondly, I am particularly grateful to my supervisor, Associate Prof. Shyamala Doraisamy and my co-supervisors in persons of Dr. Azreen Azman and Dr. Azrul Hazri Jantan for their guidance, patience, and assistance throughout this study. The support given to me by the other staff of Multimedia Department is also acknowledged.

Moreover, I would like to thank my mother and my family for their support through the period of my research. I am dedicating this study to them. I am thankful to my siblings (Dr Yusuf, Idris, Umar, Ahmad, Ibrahim, Sani) for encouraging me throughout my research. I also feel very grateful to Dr. Ibrahim Saidu, Dr. Isa Garba Abor, and Dr. Dahiru Sani for their continue support and guidance throughout the period of the study.

Finally, I wholeheartedly acknowledged the opportunity given to me by the Usmanu Danfodiyo University, Sokoto, Nigeria as well as the courage given to me by my colleagues in the University.

I certify that a Thesis Examination Committee has met on 15 September 2016 to conduct the final examination of Abubakar Roko on his thesis entitled "Effective Query Structuring with Ranking using Named Entity Categories for XML Retrieval" in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Doctor of Philosophy.

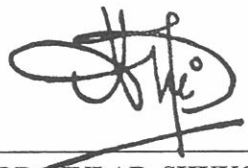
Members of the Thesis Examination Committee were as follows:

Rahmita Wirza binti O. K. Rahmat, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Muhamad Taufik bin Abdullah, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Masrah Azrifah binti Azmi Murad, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Joemon M. Jose, PhD
Professor
University of Glasgow
United Kingdom
(External Examiner)



NOR AINI AB. SHUKOR, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 3 November 2016

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

Shyamala Doraisamy, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Azreen Azman, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Azrul Hazri Jantan, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

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

Date:

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software

Signature: _____ Date: _____

Name and Matric No: Abubakar Roko (GS27173)

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) were adhered to.

Signature: _____
Name of Chairman of Supervisory Committee: Associate Professor Dr. Shyamala Doraisamy

Signature: _____
Name of Member of Supervisory Committee: Dr. Azreen Azman

Signature: _____
Name of Member of Supervisory Committee: Dr. Azrul Hazri Jantan

TABLE OF CONTENTS

		Page
ABSTARCT		i
ABSTRAK		iii
ACKNOWLEDGEMENTS		vi
APPROVAL		vii
DECLARATION		ix
LIST OF TABLES		xiv
LIST OF FIGURES		xv
LIST OF ABBREVIATIONS		xvi
CHAPTER		
1	**INTRODUCTION**	1
	1.1 Background	1
	1.2 Research Problem	2
	1.3 Motivation	3
	1.4 Research Objectives	4
	1.5 Research Scope	4
	1.6 Research Contributions	4
	1.7 Thesis Organization	5
2	**LITERATURE REVIEW**	6
	2.1 Introduction	6
	2.2 XML Overview	6
	2.3 XML Components	7
	2.3.1 Elements	7
	2.3.2 Attributes	8
	2.3.3 Comments	8
	2.4 XML Tree Structure	9
	2.5 Schema	11
	2.5.1 XML Document Type Definitions (DTDs)	12
	2.5.2 XML Schema	14
	2.5.3 Important XML DTD and XML Scheme	14
	2.6 XML Query Languages	15
	2.6.1 XML Path Language (XPath)	15
	2.6.2 XQuery	16
	2.7 XML Parsing	17
	2.7.1 Simple API for XML (SAX)	17
	2.7.2 Streaming API for XML (StAX)	18
	2.8 Related Research on Keyword Query	18
	2.8.1 LCA Based Approaches	18
	2.8.2 Statistical based approaches	19
	2.8.3 Query Structuring	22
	2.9 Conclusion	26
3	**METHODOLOGY**	27
	3.1 Introduction	27

3.2 Test Collections 28
3.2.1 Document Collections 28
3.2.2 Topics 29
3.2.3 Relevance assessments 30
3.3 Evaluation measures 31
3.3.1 Precision 32
3.3.2 Recall 32
3.3.3 Precision at k (P@k) 32
3.3.4 Mean Average Precision (MAP) 32
3.4 Retrieval Track 33
3.4.1 Ad-hoc Track 33
3.5 Notations and Retrieval Models 33
3.5.1 Notations 33
3.5.2 Retrieval Models 34
3.6 The Research Framework 37
3.6.1 Problem Formulation 38
3.6.2 Previous Systems Analysis and Implementation 38
3.6.3 Indexer 39
3.6.4 Keyword Query 41
3.6.5 The Proposed Systems 41
3.6.6 Conducting Experiments 43
3.6.7 Performance Metrics Evaluation 43
3.7 Experiments Environment 43
3.7.1 Computer Resources 43
3.7.2 Databases 43
3.7.3 Experimental Setup 44
3.7.4 Performance metrics 47
3.8 Conclusion 47

4 DOCUMENT ENRICHMENT USING SEMANTIC TAGS FOR XML RETRIEVAL 48
4.1 Introduction 48
4.2 Preliminaries 49
4.3 Document Enrichment 50
4.3.1 Semantics Tags Extraction (STSE) Method 50
4.3.2 Element Enrichment Method (EERM) 52
4.4 Performance Evaluation 54
4.4.1 Result and Discussions 55
4.5 Conclusion 55

5 KEYWORD QUERY STRUCTURING USING NEC FOR XML RETRIEVAL 56
5.1 Introduction 56
5.2 Semantic-Aware Index Scheme (SAIS) Construction 56
5.3 XML Keyword Query Structuring (XKQSS) 58
5.3.1 Entity Based Query Segmentation (EBQS) 59
5.3.2 Predicate Identification Algorithm (PIA) 60
5.3.3 Entity Identification Algorithm (EIA) 64
5.3.4 Query Formulation (QRYF) 66
5.4 Operation of the XKQSS 68

5.5 Performance Evaluation 69
5.5.1 Results and Discussions 69
5.6 Conclusion 71

6 A RANKING-AWARE XKQSS FOR XML RETRIEVAL 72
6.1 Introduction 72
6.2 Index Construction 72
6.3 Named Entity based Ranking with term proximity (rk_BM25TOPF) 73
6.3.1 Boost Score 74
6.4 Proposed RANKING-AWARE XKQSS (RAXKQSS) 78
6.4.1 Search Intention Identification (SII) 82
6.4.2 Query Formulation via Node (QRYFv) 87
6.5 Performance Evaluation 89
6.5.1 Results and Discussions 90
6.6 Conclusion 93

7 CONCLUSION AND FUTURE STUDIES 95
7.1 Conclusion 95
7.2 Future research 96

REFERENCES 97
APPENDICES 105
BIODATA OF STUDENT 108
LIST OF PUBLICATIONS 109

LIST OF TABLES

Table		Page
3.1	Abbreviations and descriptions	34
3.2	Queries on Sigmod and user target nodes	45
3.3	Queries on IMDB and user target nodes	46
3.4	Queries on IMDB and returned nodes	46
3.5	Queries on Sigmod dataset and returned nodes	47
4.1	Performance evaluation results for the three experiments	54
5.1	Leaf Nodes and their Confidence value	57
5.2	Query evaluation result on Sigmod dataset	69
5.3	Query evaluation result on IMDB dataset	69
6.1	Segmentations and Scores	83
6.2	predicate nodes and their Search Via Condition (SVC)	89
6.3	Parameter setting for rk_BM25TOPF	90
6.4	Query evaluation result on IMDB dataset	90
6.5	Query evaluation result on Sigmod dataset	91

LIST OF FIGURES

Figure		Page
2.1	An Example XML document (CoursesList.xml)	7
2.2	Tree Model of XML document	10
2.3	DTD for the XML document	12
2.4	Valid and invalid course Elements	13
3.1	Example of INEX Topic	30
3.2	Research Framework	38
3.3	Sigmod XML Element	40
3.4	Tokenized Element Tokens	40
3.5	Element tokens after stop words removal	40
3.6	Selected Index Term	41
4.1	INEX 2009 XML file 4966980.xml	49
5.1	SAIS	56
5.2	A Query and its segments	60
5.3	Three segments and two nodes	62
5.4	Generated predicates	64
5.5	Precision comparison for Sigmod	70
5.6	Precision comparison for IMDB	71
6.1	Index Structure	72
6.2	Predicates and their scores	79
6.3	Improved Predicates and their scores	80
6.4	Segments and predicates	87
6.5	Precision comparison with IMDB	92
6.6	Precision comparison with Sigmod	92

LIST OF ABBREVIATIONS

Term	Description
AST	Annotated Semantic Tag
CRNA	Compute Returned node algorithm
EBQS	Entity based query segmentation
EIA	Entity Identification Algorithm
IPIA	Improved Predicate Identification Algorithm
LCA	Lowest Common Ancestor
NBQS	N-gram based query segmentation
NEC	Named Entity Category
NER	Named Entity Recognition
PIA	Predicates Identification Algorithm
RAXKQSS	Rank Aware XML Keyword Query Structuring System
SAIS	Semantic Aware Index Scheme
SII	Search Intention Identification
SLCA	Smallest Lowest Common Ancestor
tf-idf	Term frequency - inverse document frequency
XKQSS	XML Keyword Query Structuring System

CHAPTER 1

INTRODUCTION

1.1 Background

The exponential growth of the Internet and the Web has tremendously increased the volume of information. To manage this information, Information Retrieval (IR) systems are required to extract data from the Web. These systems return a ranked list of relevant result to users based on their information needs. The information is in different formats and on a variety of subjects, which can be categorised into three groups: unstructured documents, semi-structured documents and structured documents. The unstructured documents have no fixed pre-defined format e.g. a flat-files containing textual information. In contrast to unstructured documents, the structured documents have a fixed pre-defined format, e.g. Database records. While the semi-structured documents have more tight structure than unstructured documents, but the structure is not as tight as in structured documents. Recently, structured documents and semi-structured documents are commonly represented using eXtensible Mark-up Language (XML).

XML is a mark-up language with some apparent similarities to HyperText Markup Language (HTML), but XML imposes a more rigorous structure than HTML. Unlike HTML, XML allows user-defined tags which are used to specify meaning to the data contained between them. The XML has been proposed by World Wide Web Consortium (W3C) in 1990 to specify the logical or tree structure of documents, in which separate document parts (e.g. section, paragraph, chapter, article) and their logical structure (e.g. a section and its title, a chapter made of sections) are clearly marked-up. As more and more documents are being represented in XML format, methods to retrieve them are required.

Traditionally, a query language such as XPath or XQuery is used to compose queries called structured queries to access information from documents formatted with XML (Lian, Mamoulis, Cheung, & Yiu, 2005; K. Nguyen & Cao, 2012). These queries retrieve precise answers because the queries contain specifications of what to return and where in the document to find it (Y. Li et al., 2004). However, using query languages to express queries prove to be difficult for most users since writing them requires learning a query language and knowledge of the underlying data schema. Structural heterogeneity of XML data also complicates the ability to express a query using a query language. For example, the *author* or *au* tags are often used to tag the author of a book or journal paper in an XML data. But which one of these tags will be used to compose the query? Another problem with structured queries is with regard to the way the returned results are presented to the users. A structured query returns a set of results; meaning that, the results are not presented to the user in any particular order.

An alternative approach that can overcome the problems of unknown schema and/or structure heterogeneity of XML data is the keyword search. This is where the users pose queries using a set of free formatted keywords. The difficulty in expressing user query in the form of structured queries and the success of web search engine have made many users to be familiar with keyword search (Lou, Li, & Chen, 2012; Ya-hui Chang, Cheng-Yi Wu, 2012). Consequently, users prefer to use a keyword search query interface to search XML data. Keyword search requires a user not to learn a query language and/or know the schema of the underlying documents. Although, keyword query can be issued easily, it has an inexpressive side-effect, i.e. ambiguity of expressing user's search intention (X. Lin, Wang, Xu, & Zeng, 2010). The ambiguity of the keyword query may cause large number of results to be returned and thus makes keyword query not effective (He, Wu, Chen, Zhou, & Chen, 2013; Li, X., Li, Z., Wang, P., & Chen, Q., 2010). Therefore, search systems that enhance the effectiveness of keyword queries are highly needed (X. Liu, Wan, & Chen, 2011; Zhong, Minjuan, 2012).

Several systems are proposed to improve the keyword search. These systems are categorized into three: *Lowest Common Ancestor* (LCA) *based* systems (Cohen, S., Mamou, J., Kanza, Y., & Sagiv, 2003; G. Li et al., 2007; Xu & Papakonstantinou, 2005), q*uery structuring* (Da C. Hummel, Da Silva, Moro, & Laender, 2011; Gan & Phang, 2014; J. Li, Liu, Zhou, & Ning, 2009; X. Li, Li, Wang, & Chen, 2010; Petkova, Croft, & Diao, 2009) and *statistical based* approaches (Bao, Lu, Ling, & Chen, 2010; Bao, Lu, Ling, Xu, & Wu, 2010; G. Li, Li, Feng, & Zhou, 2009; Lou et al., 2012). Among these systems, Query Structuring is the most effective because it combines the capability of user friendly nature of keyword search with effectiveness of XML query languages (e.g. XQuery). Also, (Tahraoui M. A. et al., 2013), averred that structured queries contain structured constraints which help reduce the size of irrelevant results and limit the search to document components that are relevant.

1.2 Research problem

Numerous studies have been conducted on effective query structuring for XML retrieval systems (Petkova et al., 2009, J. Li et al., 2009, Da C. Hummel et al., 2011). These studies focus on query representation, user search intention identification and ranking algorithms to improve keyword search. Despite these studies, several challenges are left unresolved in the main research issue of query structuring, which include:

Existing query structuring systems are ineffective because the systems fail to put query keyword ambiguities into consideration. The system (Petkova et al., 2009) groups the query keywords into the query keywords that match tag name of an element and the query keywords that match the data value of an element. However, the system fails to classify a query keyword that appears both as a tag name and data value of an element. The StruX system (Da C. Hummel et al., 2011) also split a query into sequence of segments, where each segment consists of a query keyword or a sequence of keywords. In this case, the system ignores to identify that a query keyword can appear in different parts of an XML document having different semantics because it

considers a query just as a sequence of keywords not as a sequence of semantically related terms.

The StruX system (Da C. Hummel et al., 2011) is impractical to be used in an environment where DTD is not part of the XML document because entity nodes in the target XML document are computed based on the heuristics apply on the document DTD. Also, the heuristics classified multi-valued attributes and grouping nodes as entity nodes and cannot handle the situation where a node's tag name belong to three classes of nodes: **-node* node, *connection* node, and *attribute* node at the same time. These lead to results that is not informative.

The systems select structured queries that fail to match user's input query because their scoring functions ignore to consider the semantics relationship between context node and predicate nodes.

Furthermore, the systems' result ranking schemes are often expressed as functions based on the tightness of the XML elements and *tf-idf* score (Petkova et al., 2009, J. Li et al., 2009). However, these schemes ignore the semantics of XML documents and therefore the systems return misleading results because the functions are powerless in recognizing irrelevant results when XML fragments have highest term frequencies.

1.3 Motivation

The exponential growth of XML documents on the Web poses the challenge of how best these documents can be accessed. The effective search approach (i.e. structured query) to search these data proves to be difficult for most users (J. Li et al., 2009) while the keyword search approach proves to be ineffective (X. Lin et al., 2010). Therefore, an effective XML keyword search engine is needed that addresses the following challenges.

1. Resolves keyword ambiguity problem: (i) a query keyword can appear as a tag name as well as a data value of some nodes within an XML document. (ii) a query keyword can appear in different XML elements having different meanings in each element.
2. Identifies the user search intention, which includes finding the target XML nodes users are interested (i.e. returned nodes) and predicates.
3. Generates and selects the best generated structured queries that represent the user's original intention.
4. Ranking functions are needed to score each relevant XML element returned by the system.

Unfortunately, existing systems cannot thoroughly resolve these challenges and hence the outputs returned by the systems are still far from being satisfactory from user's perspective.

1.4 Research Objectives

Main Goal: To design and evaluate a query structuring system for XML Retrieval. Specific goals:

1. To propose a document enrichment method to investigate its effect on keyword queries.
2. To propose a query pre-processing methods to resolve keyword query ambiguities.
3. To propose algorithms to identify user search intention from a keyword query. Specifically, this includes an algorithm to identify predicates and an algorithm to identify the XML nodes users are searching for in order to make the system return informative result.
4. To propose a field weight ranking function, which includes not only the tf-*idf* score but also the weight of a query term based on its position and the query terms proximity in order to reduce the impact of terms with high frequencies.

1.5 Research Scope

This study mainly focuses on how to create an effective keyword query structuring system for the retrieval of documents formatted using XML where the system does not extend to the environment in which documents need regular updating because static labelling scheme is used to assign unique label to each element node in XML documents. The system receives a user queries formatted as a list of keywords not natural language, which denotes user's information need. The items retrieved by the system are either a documents or elements. The system also includes a ranking function, query segmentation, user search intention identification, and query formulation.

1.6 Research Contributions

1. A document enrichment method is proposed to investigate its impact on keyword queries using Semantic Tags Extraction method (STSE) and Element Enrichment Method (EERM).
2. A Semantic Aware Indexing Scheme (SAIS) is proposed to store the Named Entity Category (NEC) of each indexed term, in addition to the usual information such as term frequencies, term position, as well as an element that contains the term in the inverted index.
3. An Entity Based Query Structuring (EBQS) method is proposed based on NECs to re-express the user's query as a list of keywords and named entities.
4. An N-gram Based Query Structuring (NBQS) method is proposed based on N-grams to interpret user input query as a list of semantic units to help address keyword query ambiguities.
5. A Predicate Identification Algorithm (PIA) is proposed to compute relevant predicates.
6. An Improved Predicate Identification Algorithm (IPIA) is proposed to compute relevant predicates.

7. A Compute Return Node Algorithm (CRNA) is proposed to returns informative result.
8. An Entity Identification Algorithm (EIA) algorithm is proposed to compute entity nodes from XML data which makes a robust system.
9. A ranking function, *rk*_BM25TOPF, is proposed to mitigate the effect of terms with high frequencies.
10. A Query Formulation (QRYF) algorithm to select the best queries which reflect user search intention.

1.7 Thesis Organization

The rest of this study is organized as follows:

Chapter 2: Presents a descriptive illustration of the basic concepts of XML and discusses the related works on existing keyword search systems.

Chapter 3: Presents the methodology adopted in this study and the proposed framework.

Chapter 4: This chapter describes the preliminary study conducted to investigate the effect of using semantic tag as part of corresponding element text content for keyword search. The Chapter also presents the performance evaluation.

Chapter 5: Describes XKQSS, an XML keyword query structuring system for XML retrieval. The chapter also presents the experiments and compares the performance of XKQSS with the XReal and the StruX.

Chapter 6: Describes a Ranking Aware XKQSS (RAXKQSS) for XML retrieval. The chapter also presents the experiments and compares the performance of RAXKQSS with StruX and SLCA.

Chapter 7: Concludes the thesis and list several future research directions on the topic of keyword query structuring.

REFERENCES

Akritidis, L., Katsaros, D., & Bozanis, P. (2012). Improved retrieval effectiveness by efficient combination of term proximity and zone scoring: A simulation-based evaluation. *Simulation Modelling Practice and Theory*, *22*, 74–91. http://doi.org/10.1016/j.simpat.2011.12.002

Almelibari, A. (2015). *Labelling Dynamic XML Documents: A GroupBased Approach* (phdthesis). University of Sheffield.

Anders Berglund , Scott Boag , Don Chamberlin , Mary F. Fernández, Michael Kay , Jonathan Robie, J. (2010). XML Path Language (XPath) 2.0 (Second Edition). Retrieved from https://www.w3.org/TR/xpath20/

Anjali, Jivani, G., & Anjali, M. (2007). A Comparative Study of Stemming Algorithms. *October*, *2*(2004), 1930–1938. http://doi.org/10.1.1.642.7100

Armstrong, E., Bodoff, S., Carson, D., Fisher, M., Green, D., & Haase, K. (2002). *The Java Web Services Tutorial*. book, Addison-Wesley.

Bao, Z., Lu, J., Ling, T. W., & Chen, B. (2010). Towards an Effective XML Keyword Search. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, *22*(8), 1077–1092.

Bao, Z., Lu, J., Ling, T. W., Xu, L., & Wu, H. (2010). An effective object-level XML keyword search. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5981 LNCS*(PART 1), 93–109. http://doi.org/10.1007/978-3-642-12026-8_10

Bao, Z., Zeng, Y., Ling, T. W., Zhang, D., Li, G., & Jagadish, H. V. (2015). A general framework to resolve the MisMatch problem in XML keyword search. *The VLDB Journal*, 1–26. http://doi.org/10.1007/s00778-015-0386-1

Berger, A., & Lafferty, J. (1999). Information retrieval as statistical translation. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '99*, 222–229. http://doi.org/10.1145/312624.312681

Brett D. McLaughlin, J. E. (2007). *Java & XML: Solutions to Real World Problems*.

Brian, D. (2006). *The Definitive Guide to Berkeley DB XML*. (N. Sixsmith, Ed.). New York, New York, USA: Apress.

Bruce W. Croft, Donald Metzler, T. S. (2010). *Search Engines Information Retrieval in Practice* (Vol. 1). book, Addison Wesley.

Büttcher, S., Clarke, C. L. A., & Cormack, G. V. (2010). *Information retrieval: Implementing and evaluating search engines*. book, Mit Press.

Büttcher, S., Clarke, C. L. a., & Lushman, B. (2006). Term proximity scoring for ad-hoc retrieval on very large text collections. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06* (p. 621). http://doi.org/10.1145/1148170.1148285

Calado, P., da Silva, A. S., Vieira, R. C., Laender, A. H. F., & Ribeiro-Neto, B. a. (2002). Searching web databases by structuring keyword-based queries. *Proceedings of the Eleventh International Conference on Information and Knowledge Management - CIKM '02*, 26. http://doi.org/10.1145/584800.584801

Clark, J., DeRose, S., & others. (1999). XML path language (XPath) version 1.0 [misc].

Cleverdon, C. (1967). The Cranfield Tests on Index Language Devices. In *Aslib Proceedings* (Vol. 19, pp. 173–194). http://doi.org/ 10.1108/eb050097

Clough, P., & Sanderson, M. (2013). Evaluating the performance of information retrieval systems using test collections. *Information Research*, *18*(2). article.

Clough, P., & Sanderson, M. (2013). Evaluating the performance of information retrieval systems using test collections. *Information Research*, *18*(2), 1–15. Retrieved from http://informationr.net/ir/18-2/paper582.html

Cohen, S., Mamou, J., Kanza, Y., & Sagiv, Y. (2003). XSEarch : A Semantic Search Engine for XML. In *Proceedings of the 29th international conference on Very large data bases*.

Connolly, T. M., & Begg, C. E. (2005). *Database systems: a practical approach to design, implementation, and management*. book, Pearson Education.

Da C. Hummel, F., Da Silva, A. S., Moro, M. M., & Laender, A. H. F. (2011). Automatically generating structured queries in XML keyword search. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 6932 LNCS, pp. 194–205). http://doi.org/10.1007/978-3-642-23577-1_17

Feng, J., Li, G., Wang, J., & Zhou, L. (2010). Finding and ranking compact connected trees for effective keyword proximity search in XML documents. *Information Systems*, *35*(2), 186–203. http://doi.org/10.1016/j.is.2009.05.004

Fuhr, N., Lalmas, M., & Kazai, G. (n.d.). INEX: Initiative for the Evaluation of XML retrieval. 2002. In *University of Dortmund*. article.

Gan, K. H., & Phang, K. K. (2014). A query transformation framework for automated structured query construction in structured retrieval environment. *Journal of Information Science*, *40*(2), 249–263. http://doi.org/ 10.1177/ 0165551513519240

Géry, M., & Largeron, C. (2012). BM25t: A BM25 extension for focused information retrieval. *Knowledge and Information Systems*, *32*(1), 217–241. http://doi.org/10.1007/s10115-011-0426-0

Géry, M., Largeron, C., & Thollard, F. (2008). Integrating structure in the probabilistic model for Information Retrieval. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on* (pp. 763–769). Retrieved from http://hal-ujm.ccsd.cnrs.fr/ujm-00331482/

Hagen, M., Potthast, M., Stein, B., & Braeutigam, C. (2010). The power of naive query segmentation. *Annual ACM Conference on Research and Development in Information Retrieval*, 797–798. http://doi.org/10.1145/1835449.1835621

Harold, E. R. (2003). *Processing XML with Java: A Guide to SAX, DOM, JDOM, JAXP, and TrAX*. Addison-Wesley.

He, T., Wu, G., Chen, Q., Zhou, J., & Chen, Z. (2013). Effectively return query results for keyword search on XML data. In *Web-Age Information Management* (pp. 802–805).

Hiemstra, D. (1998). A Linguistically Motivated Probabilistic Model of Information Retrieval. *Research and Advanced Technology for Digital Libraries*, 515–515. http://doi.org/10.1007/3-540-49653-X_34

Holzner, S. (2003). *XPath Kick Start: Navigating XML with XPath 1.0 and 2.0*. book, Sams.

Jervidalo, J. (2010). *Improving Performance of Biomedical Information Retrieval using Document- Level Field Boosting and BM25F Weighting*.

Jiang, J., Deng, Z., Gao, N., & Lv, S. (2012). Guess What I Want : Inferring the Semantics of Keyword Queries Using Evidence Theory. In *Web Technologies and Applications* (pp. 388–398).

Kim, J., Xue, X., & Croft, W. B. (2009). A probabilistic retrieval model for semistructured data. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5478 LNCS*, 228–239. http://doi.org/10.1007/978-3-642-00958-7_22

Koloniari, G., & Pitoura, E. (2010). LCA-based selection for XML document collections. *Proceedings of the 19th International Conference on World Wide Web - WWW '10*, 511. http://doi.org/10.1145/1772690.1772743

Lalmas, M. (2009). *XML RETRIEVAL*. (G. Marchionini, Ed.). Morgan & Claypool. http://doi.org/10.2200/S00203ED1V01Y200907ICR007

Lalmas, M., Rolleke, T., Szlavik, Z., & Tombros, T. (2004). Accessing XML documents: the INEX initiative. *DELOS WP7 Workshop on the Evaluation of Digital Libraries*. Retrieved from http://www.dcs.qmul.ac.uk/~mounia/CV/Papers/inextracks.pdf

Lassila, M., Junkkari, M., & Kekalainen, J. (2015). Comparison of two XML query languages from the perspective of learners. *Journal of Information Science*, *41*(5), 584–595. http://doi.org/10.1177/0165551515585259

Lavrenko, V., & Croft, W. B. (2001). Relevance-Based Language Models. *Sigir'01*, 8. http://doi.org/10.1145/383952.383972

Li, G., Li, C., Feng, J., & Zhou, L. (2009). SAIL: Structure-aware indexing for effective and progressive top-k keyword search over XML documents. *Information Sciences*, *179*(21), 3745–3762. http://doi.org/10.1016/j.ins.2009.06.025

Li, G., Li, G., Feng, J., Feng, J., Wang, J., Wang, J., … Zhou, L. (2007). Effective keyword search for valuable lcas over xml documents. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 31–40). http://doi.org/10.1145/1321440.1321447

Li, J., Liu, C., Zhou, R., & Ning, B. (2009). Processing XML Keyword Search by Constructing Effective Structured Queries. *Advances in Data and Web Management*.

Li, J., Liu, C., Zhou, R., & Wang, W. (2010). Suggestion of promising result types for XML keyword search. In *Proceedings of the 13th International Conference on Extending Database Technology - EDBT '10* (pp. 561–572). New York, New York, USA: ACM Press. http://doi.org/10.1145/1739041.1739108

Li, X., Li, Z., Chen, Q., & Li, N. (2011). XIOTR :A Terse Ranking of XIO for XML Keyword Search. *Journal of Software*, *6*(1), 156–163. http://doi.org/10.4304/jsw.6.1.156-163

Li, X., Li, Z., Wang, P., & Chen, Q. (2010). XIOF: Finding XIO for Effective Keyword Search in XML Documents. In *2010 2nd International Workshop on Intelligent Systems and Applications* (pp. 1–6). Ieee. http://doi.org/10.1109/IWISA.2010.5473249

Li, Y., Li, Y., Yu, C., Yu, C., Jagadish, H. V., & Jagadish, H. V. (2004). Schema-free XQuery. In *Very Large Data Bases* (pp. 72–83). http://doi.org/DOI: 10.1016/B978-012088469-8.50010-3

Lian, W., Mamoulis, N., Cheung, D. W. L., & Yiu, S. M. (2005). Indexing useful structural patterns for XML query processing. *IEEE Transactions on Knowledge and Data Engineering*, *17*(7), 997–1009. http://doi.org/10.1109/TKDE.2005.110

Lin, X., Wang, N., Xu, D., & Zeng, X. (2010). A novel XML keyword query approach using entity subtree. *Journal of Systems and Software*, *83*(6), 990–1003. http://doi.org/10.1016/j.jss.2009.12.024

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google Books Ngram Corpus. *Proceedings of the ACL 2012 System Demonstrations*, (July), 169–174. Retrieved from http://www.aclweb.org/anthology/P12-3029

Liu, D., Wan, C., Chen, L., Liu, X., & Nie, J.-Y. (2013). Weighting tags and paths in XML documents according to their topic generalization. *Information Sciences*, *249*, 48–66. http://doi.org/10.1016/j.ins.2013.06.019

Liu, X., Wan, C., & Chen, L. (2011). Returning clustered results for keyword search on XML documents. *IEEE Transactions on Knowledge and Data Engineering*, *23*(12), 1811–1825. http://doi.org/10.1109/TKDE.2011.183

Liu, Z., & Chen, Y. (2007). Identifying meaningful return information for XML keyword search. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data - SIGMOD '07* (p. 329). New York, New York, USA: ACM Press. http://doi.org/10.1145/1247480.1247518

Lopez-Veyna, J. I., Sosa-Sosa, V. J., & Lopez-Arevalo, I. (2012). KESOSD: keyword search over structured data. In *Proceedings of the Third International Workshop on Keyword Search on Structured Data* (Vol. 1, pp. 23–31). http://doi.org/10.1145/2254736.2254743

Lou, Y., Li, Z., & Chen, Q. (2012). Semantic relevance ranking for XML keyword search. *Information Sciences*, *190*, 127–143. http://doi.org/10.1016/j.ins.2011.12.011

Lou, Y., Wu, Q., Ji, B., Zheng, R., Zhang, M., & Wei, W. (2013). Effective Entity Unit for XML Keyword Search ⋆. *Computational Information Systems*, *17*(9), 6811–6818. http://doi.org/10.12733/jcis6510

Lu, W., Robertson, S., & MacFarlane, A. (2006). Field-Weighted XML Retrieval Based on BM25. In *Advances in XML Information Retrieval and Evaluation* (Vol. 3977, pp. 161–171). http://doi.org/10.1007/11766278

Manning, C. D., Raghavan, P., Schütze, H., & others. (2008). *Introduction to information retrieval* (Vol. 1). book, Cambridge university press Cambridge.

Møller, A., & Schwartzbach, M. I. (2006). *An introduction to XML and Web Technologies*. book, Pearson Education.

Nguyen, K., & Cao, J. (2010). Exploit Keyword Query Semantics and Structure of Data for Effective XML Keyword Search. In *Proceedings of the Twenty-First Australasian Conference on Database Technologies* (Vol. 104, pp. 133–140).

Nguyen, K., & Cao, J. (2012). Top-K data source selection for keyword queries over multiple XML data sources. *Journal of Information Science*, *38*(2), 156–175. http://doi.org/10.1177/0165551511435875

Peters, C. (2002). *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001. Revised Papers*. book, Springer Science & Business Media.

Petkova, D., Croft, W. B., & Diao, Y. (2009). Refining Keyword Queries for XML Retrieval by Combining Content and Structure. *Advances in Information Retrieval*.

Piwowarski, B., & Lalmas, M. (2004). Providing consistent and exhaustive relevance assessments for XML retrieval evaluation. *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management - CIKM '04*, 361. http://doi.org/10.1145/1031171.1031246

Piwowarski, B., Trotman, A., & Lalmas, M. (2008). Sound and complete relevance assessment for XML retrieval. *ACM Transactions on Information Systems (TOIS)*, *27*, 1:1–1:37. http://doi.org/10.1145/1416950.1416951

Ponte, J., & Croft, B. (1998). A Language Modeling Approach To Information Retrieval. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 275–81. http://doi.org/10.1145/290941.291008

Robertson, S. . (1997). The probability ranking principle in IR. *Reading in Information Retrieval*, 281–286.

Schenkel, R. (2006). Structural Feedback for Keyword Based XML Retrval. In *Lecture Notes in Computer Science* (pp. 326–337).

Schenkel, R., & Theobald, M. (2006). Structural feedback for keyword-based XML retrieval. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 3936 LNCS, pp. 326–337). http://doi.org/10.1007/11735106_29

Singh, V., & Saini, B. (2014). An Effective Pre-processing Algorithm for Information Retrieval Systems. *International Journal of Database Management Systems*, *6*(6), 13.

Softwarecave. (2014). Retrieved from https://softwarecave.org/2014/02/18/parse-xml-document-using-streaming-api-for-xml-stax/

Sparck Jones, K., Walker, S., & Robertson, S. . (2000). A Probabilistic Model of Onformation Retrieval: development and comparative experiments. *Information Processing & Management*, *36*, 809–840. http://doi.org/10.1016/S0306-4573(00)00016-9

Strohman, T., Metzler, D., Turtle, H., & Croft, W. (2005). Indri: A language model-based search engine for complex queries. *Proceedings of the International Conference on Intelligent Analysis*, *2*(6), 2--6. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.3502&rep=rep1&type=pdf

Tahraoui, M. A., Pinel-Sauvagnat, K., Laitang, C., Boughanem, M., Kheddouci, H., & Ning, L. (2013). A survey on tree matching and XML retrieval. *Computer Science Review*, *8*, 1–23. http://doi.org/10.1016/j.cosrev.2013.02.001

Tannier, X. (2005). From natural language to NEXI , an interface for INEX 2005 queries. *Springer*.

Tannier, X., & Geva, S. (2005). XML Retrieval with a Natural Language Interface. *Springer*, 29–40.

Tatarinov, I., Beyer, K., & Shekita, E. (2002). Storing and Querying Ordered XML Using a Relational Database System.

Thw Java™ Tutorials. (2015). Retrieved from https://docs.oracle.com/javase/tutorial/jaxp/sax/

Voorhees, E. M., Harman, D. K., & others. (2005). *TREC: Experiment and evaluation in information retrieval* (Vol. 1). book, MIT press Cambridge.

Wang, Q., Li, Q., Wang, S., & Du, X. (2010). Exploiting semantic tags in XML retrieval. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6203 LNCS*, 133–144. http://doi.org/10.1007/978-3-642-14556-8_15

Wang, Q., Ramírez, G., Marx, M., Theobald, M., & Kamps, J. (2011). Overview of the INEX 2011 Data-Centric Track. *Focused Retrieval of Content and Structure - Lecture Notes in Computer Science, Vol. 7424*, 118–137.

Woodley, A., & Geva, S. (2006). NLPX at INEX 2005. *Lecture Notes in Computer Science*, (3977), 358–372.

Xu, Y., & Papakonstantinou, Y. (2005). Efficient keyword search for smallest LCAs in XML databases. In *International Conference on Management of Data*. http://doi.org/10.1145/1066157.1066217

Ya-hui Chang, Cheng-Yi Wu, C.-C. Lo. (2012). Processing XML Queries with Structural and Full-Text Constraints. *Information Science and Engineering*, *242*, 221–242.

Zhai, Chengxiang, S. M. (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining.* ACM and Morgan & Claypool Publishers.

Zhai, C., & Lafferty, J. (2004). A Study of Smoothing Methods for Language Models Applied to Information Retrieval, *22*(2), 0–33.

Zhi-xian, T., Jun, F., Li-ming, X., & Ya-qing, S. (2015). A Bottom-up Algorithm for XML Twig Queries. *International Journal of Database Theory and Application*, *8*(4), 49–58.

Zhong, M. (2012). Selecting Good Expansion Terms for Improving XML Retrieval Performance. *2012 International Conference on Control Engineering and Communication Technology*, 480–483. http://doi.org/10.1109/ICCECT.2012.176