

**Deanship of Graduate Studies
Al-Quds University**



**System for Top-k Keyword Search processing over
relational databases using semantics**

Samia Taha Hussein Abdulhay

M.SC. Thesis

Jerusalem – Palestine

1431 / 2010

**System for Top-k Keyword Search processing over
relational databases using semantics**

**Prepared By:
Samia Taha Hussein Abdulhay**

**B.Sc. : Computer Science, Al-Quds University,
Palestine**

**Supervisor :
Dr. Rashid Jayousi**

**A thesis Submitted in Partial Fulfillment of
Requirements for the Master Degree of Computer
Science from Computer Science Department of Al-
Quds University.**

1431 / 2010

**Deanship of Graduate Studies
Al-Quds University**



Thesis Approval

**System for Top-k Keyword Search processing over relational
databases using semantics**

**Prepared By: Samia Taha Hussein Abdulhay
Registration No: 20510194**

Supervisor: Dr. Rashid Jayousi

Master thesis submitted and accepted Date: 15/5/2010

**The names and signatures of examining committee members are
follows:**

- | | |
|---|------------------------|
| 1. Head of Committee: Dr. Rashid Jayousi | Signature:..... |
| 2. Internal Examiner: Dr. Nedal Kafri | Signature:..... |
| 3. External Examiner: Dr. Yousef Abuzir | Signature:..... |

Jerusalem – Palestine

1431/ 2010

Dedication

I would like to dedicate this work to my family, who supported me in all phases of this thesis, especially to my lovely mother, for her support, care and love, and to all the people who help me to overcome difficulties.

Samia taha Hussein Abdulhay

Declaration

I certify that this thesis submitted for the degree of Master of Computer Science, is the result of my own research, except where otherwise acknowledged, and that this thesis (or any part of the same) has not been submitted for a higher degree to any other university or institution.

Signed

Samia Taha Hussein Abdulhay

Date: 15/5 /2010

Acknowledgements

Grateful thanks for God for giving me the patience and power to complete this work.

I sincerely thank Al-Quds University for giving me the opportunity to study the M.Sc., and its efforts and help throughout this study.

I am deeply grateful, despite the inability of these words to express my thanks to my supervisor, Dr. Rashid Jayousi, for his fruitful discussion, a valuable guidance, continuous support, kindness, and allowing me a lot of his time.

Special thanks to the Library department of Al-Quds University for allowing me to use the Library database to conduct the experiments.

Last, my endless thanks to my mother for her never ending enthusiasm, and to my family for their encouragement and support.

Finally, to the soul of my father.....

Abstract

A variety of keyword search systems over relational database are widely known. Generally such systems do not take into account what is meant from the search query, the user type a list of keywords in the search query, and then the search system retrieve all the results (a set of related records) that contains this keywords, which leads to a high irrelevant results at first top-k. In order to improve the relevancy of such results, this thesis proposed a simple keyword search technique that can help ordinary users to be more specific in expressing their needs.

This can be done by adding some schema information (e.g., table name, field name), which can be used as semantics to the searching keywords. This thesis presents *Ssearch* system that is designed to handle the proposed idea.

The researcher has conducted several experiments that use the Library database of Al-Quds University. The experimental results showed that *Ssearch* adds a significant improvement in terms of relevancy with acceptable overhead time when compare it with an existing approach.

المخلص

يوجد اليوم عدة أنظمة تدعم عملية البحث باستخدام الكلمات المفتاحية (Keyword Search)

على قواعد البيانات العلائقية، وقد انتشرت بشكل واسع، وذلك لسهولة استخدامها من قبل المستخدم العادي، ولكن عملية البحث باستخدام الكلمات المفتاحية لا تأخذ بعين الاعتبار المعنى الذي يقصده المستخدم من عملية البحث، مما يتسبب ذلك في ارجاع عدد كبير من النتائج في الصفحات الاولى ليست ذات علاقة بما يقصده المستخدم، على سبيل المثال، لو اراد المستخدم ان يبحث عن الكتب التي قام بتأليفها المؤلف طه حسين فان عملية البحث ممكن ان تعطيك ليس فقط الكتب التي الفها المؤلف طه حسين فحسب وانما الكتب التي تتحدث عن طه حسين وبيانات المستخدمين الذين يحملون نفس الاسم ايضا، لذلك اقترحت الباحثة في هذه الاطروحة فكرة جديدة لتحسين دقة النتائج التي تظهر في الصفحات الاولى، وذلك باعطاء المستخدم مرونة اكثر للتعبير عن احتياجاته بشكل ادق وذلك بالاستفادة من بعض المعلومات المتوفرة في بنية نظام قواعد البيانات مثل اسم الجدول او اسم الحقل لاستخدامه كمعنى للكلمات المفتاحية التي يستخدمها المستخدم في البحث، وبالتالي مساعدة نظام الحاسوب على تمييز النتائج التي لها صلة بما يعنيه المستخدم بشكل اكبر. لذلك تم اقتراح لغة استعلام بسيطة تنفيذ من هذه المعلومات لتحسين دقة النتائج التي تظهر اولاً.

لقد تم اجراء عدة تجارب باستخدام قاعدة بيانات مكتبة جامعة القدس، وقد اثبتت هذه التجارب تحسنا ملحوظا في دقة نتائج البحث مع زيادة طفيفة ومقبولة في الوقت اللازم لبدء اخراج هذه النتائج للمستخدم في النظام المقترح *Ssearch* بالمقارنة مع نظام اخر قائم.

Table of Contents

Declaration.....	i
Acknowledgements.....	ii
Abstract.....	iii
المخلص.....	iv
Table of Contents.....	v
List of Figures.....	viii
List of Tables.....	x
Chapter One: Introduction.....	1
1.1 Introduction.....	1
1.2 History of Keyword search.....	1
1.3 Related Work.....	2
1.4 Problems and Objectives.....	3
1.5 Motivation for this thesis.....	3
1.6 Contribution in the field.....	4
Chapter Two: Literature Review.....	5
2.1 Introduction.....	5
2.2 Graph Model of Relational Databases.....	5
2.3 BANKS.....	6
2.3.1 Database and query model.....	6
2.3.2 Backward expanding search algorithm.....	7
2.3.3 User Feedback.....	8
2.4 DBXplorer.....	9
2.4.1 Publish procedure.....	9
2.4.2 Search procedure.....	10
2.4.3 Execution efficiency.....	11
2.5 DISCOVER.....	11
2.5.1 Architecture.....	12
2.6 Efficient IR style keyword search.....	14
2.6.1 System Architecture.....	14
2.7 ObjectRank.....	16
2.7.1 ObjectRank Architecture.....	17
2.8 Comparisons of different techniques.....	18
Chapter Three: Background.....	19

3.1 Relational Database	19
3.2 Extracting data from database.....	20
3.2.1 SQL	20
3.2.2 QBE.....	21
3.2.3 Keyword Search.....	22
3.2.4 Subject Search.....	24
3.2.5 SQL Vs. QBE.....	25
3.2.6 Keyword search vs. SQL	25
3.2.7 Keyword vs. Subject Search	26
3.3 Stemming	27
3.4 Stop Words.....	27
3.5 Graph.....	29
3.5.1 Undirected graphs representation	30
3.6 Dijkstra's algorithm	32
3.7 Quicksort algorithm	33
Chapter Four: Discover Model	35
4.1 Introduction.....	35
4.2 Discover Architecture	37
4.3 Search Query.....	38
4.4 Master Index	38
4.5 Candidate Networks Generator.....	39
4.6 Execution Plan	39
4.6.1 Ranking Algorithm	40
4.7 Characteristics of Discover model.....	41
4.8 Summery.....	41
Chapter Five: Ssearch Model.....	42
5.1 Introduction.....	42
5.2 Ssearch Architecture	44
5.3 Offline System	45
5.3.1 Offline System Algorithms	49
5.3.2 Updating the Master Index Database	53
5.4 Online System.....	54
5.4.1 Suggested query syntax.....	54
5.4.2 Parsing the Query String Algorithm	56
5.4.3 Retrieving Matching Records Algorithm.....	58
5.4.4 Semantic Satisfaction Algorithm	59
5.4.5 Tuples Combinations	61

5.4.6 Data Graph	63
5.4.7 Candidate Network Generator Algorithm.....	65
5.4.8 Pruning Candidate Networks Algorithm	67
5.4.9 Ranking Algorithm	68
5.4.10 SQL Answers	69
5.5 Characteristics of <i>Ssearch</i>	79
5.6 Summery	79
Chapter Six: Experimental Design and Results Analysis	80
6.1 Introduction.....	80
6.2 Experiments Setup	80
6.3 Experiments Metrics	81
6.4 Experiments Outline	83
6.4.2 Experiment 2: Testing the scalability of the system in terms of relevancy and overhead time	91
Chapter 7: Conclusion and Future Work	105
7.1 Conclusion	105
7.2 Future Work.....	107
References.....	108
List of Appendixes.....	110
Appendix A: Experimental Query Set	110
Appendix B: Term Index	111

List of Figures

Figure 2.1: Discover Architecture.....	13
Figure 2.2: IR style keyword search architecture.	16
Figure 3.1: ER- Diagram of the Library database.....	20
Figure 3.2: QBE interface for Ms Access.....	22
Figure 3.3: Keyword search interface.....	23
Figure 3.4: Subject search interface.....	24
Figure 3.5: Example for undirected graph representation using adjacency matrix	30
Figure 3.6: Example for undirected graph representation using adjacency matrix.	31
Figure 3.7: Dijkstra's Algorithm.....	33
Figure 3.8: Quicksort Algorithm.....	34
Figure 4.1: Keyword search results.....	35
Figure 4.2: TPC-H schema [Hristidis, 2002].....	36
Figure 4.3: Instance of TPC-H schema.....	36
Figure 4.4: Discover Architecture.....	37
Figure 5.1: Library Schema	43
Figure 5.2: Sample of Library Instance.....	43
Figure 5.3: Ssearch Architecture.....	44
Figure 5.4: The local copy of the Original DB.....	50
Figure 5.5: Output of the Parsing Algorithm (Parsing Matrix).....	58
Figure 5.6: The Output of Retrieving Matching Records Algorithm.....	59
Figure 5.7: The Output of the Semantic Satisfaction Algorithm.....	61
Figure 5.8: The output of the Tuple-Combinations Generator Algorithm (TuplePairs matrix).....	62
Figure 5.9: Graph of the database instance in Figure (5.2).....	63
Figure 5.10: The output of the Data Graph representation using Adjacency List.....	64
Figure 5.11: Output of the candidate network generator algorithm.....	66
Figure 5.12: The candidate networks after pruning.....	67
Figure 5.13: The order of the joining networks after ranking and their corresponding ranking scores.....	69
Figure 5.14: Example of the needed information for generating an equivalent SQL statement for the joining network $3 \infty 14 \infty 7 \infty 1 \infty 12$	72
Figure 5.15: The different parts of the equivalent SQL statement for the joining network $3 \infty 14 \infty 7 \infty 1 \infty 12$	77
Figure 6.1: ER-Diagram of the experimental DB.....	81
Figure 6.2: Some relevant and irrelevant answers.....	82
Figure 6.3: Relevancy at Top-10 (10 MB).....	85
Figure 6.4: Relevancy at Top-20 (10 MB).....	86
Figure 6.5: Relevancy at Top-30 (10 MB).....	87
Figure 6.6: Relevancy at Top-40 (10 MB).....	88
Figure 6.7: Relevancy at Top-50 (10 MB).....	89
Figure 6.8: Mean precision of <i>Discover</i> and <i>Ssearch</i> (10 MB).....	91
Figure 6.9: Relevancy at Top-10 (29 MB).....	93
Figure 6.10: Relevancy at Top-20 (29 MB).....	94
Figure 6.11: Relevancy at Top-30 (29 MB).....	95
Figure 6.12: Relevancy at Top-40 (29 MB).....	96
Figure 6.13: Relevancy at Top-50 (29 MB).....	97

Figure 6.14: Mean precision of Discover and Ssearch (29 MB).98
Figure 6.15: The relation between number of matching keywords and the overhead
fraction per query..... 101

List of Tables

Table 3.1: Differences between SQL and QBE.....	25
Table 3.2: Differences between keyword search and SQL.....	26
Table 3.3: Differences between keyword search and subject search.....	26
Table 3.4-a: Stopwords [www.wenconfs.com].....	27
Table 5.1: Keywords table of the Library DB.	46
Table 5.2: Keywords Information table of the Library DB.	47
Table 5.3: Tuples Information table of the Library DB.....	48
Table 5.4: Primary to Foreign key table of the Library DB.....	49
Table 5.5: Examples of the suggested query syntax.....	56
Table 5.6: Pairs of tuples that have primary to foreign key relationship.....	65
Table 5.7: The candidates networks and their equivalent SQL statements of the query “author:nancy book:planning”	78
Table 6.1: Semantics Overheads per query.....	101

Chapter One: Introduction

1.1 Introduction

Keyword search querying has emerged as one of the most effective paradigms for information discovery, especially over relational database. One of the key advantages of keyword search querying is its simplicity – users do not have to learn a complex query language, and can issue queries without any prior knowledge about the structure of the underlying data. Since the keyword search query interface is very flexible, queries may not always be precise and can potentially return a large number of query results, especially in a large amount of data stored within the database. Consequently an important requirement for keyword search is to rank the query results so that the most relevant results appear first.

This thesis proposes a new system where the semantics are a part of the search query which allows the user to be more precise in formulating the query to improve the relevancy of the query results.

This thesis starts by illustrating the previous related work in chapter 2. Chapter 3 provides the needed background about different information retrieval approaches over databases. An existing Approach (*Discover*) is presented in chapter 4. Then in chapter 5 the proposed system (*Ssearch*) is described in detail. The experimental design and results analysis are in chapter 6. Finally, conclusion and future work are in chapter 7.

1.2 History of Keyword search

Keyword search appears first as a tool for information retrieval over internet where the search engines provide keyword search on top of sets of documents, when a set of keywords is provided by the user, the search engine returns all documents that are associated with these keywords.

Excite introduced the concepts of keyword searching over internet, it was launched in February 1993 by Stanford students and was then called Architext. They had the idea of using statistical analysis of word relationships to make searching more efficient. They were