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** **Signiture:……..**
2. **Internal Examiner: Dr. Nedal Kafri** **Signiture:……..**
3. **External Examiner: Dr. Yousef Abuzir** **Signiture:……..**

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.

الملخص

يوجد اليوم عدة انظمة 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 لاسخدامه 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 تستفيد من هذه المعلومات لتحسين دقة النتائج التي تظهر اولا.

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

List of Figures

List of Tables

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