Keyword Search in Databases

Synthesis Lectures on Data Management

Editor
M. Tamer Özsu

Keyword Search in Databases
Jeffrey Xu Yu, Lu Qin, and Lijun Chang
2010

Keyword Search in Databases

Jeffrey Xu Yu, Lu Qin, and Lijun Chang

ISBN: 978-3-031-79425-4 paperback
ISBN: 978-3-031-79426-1 ebook

DOI 10.1007/978-3-031-79426-1

A Publication in the Springer series
SYNTHESIS LECTURES ON DATA MANAGEMENT

Lecture #1
Series Editor: M. Tamer Özsu
Series ISSN
Synthesis Lectures on Data Management
ISSN pending.

Keyword Search in Databases

Jeffrey Xu Yu, Lu Qin, and Lijun Chang
Chinese University of Hong Kong

SYNTHESIS LECTURES ON DATA MANAGEMENT #1

ABSTRACT

It has become highly desirable to provide users with flexible ways to query/search information over databases as simple as keyword search like Google search.

This book surveys the recent developments on keyword search over databases, and focuses on finding structural information among objects in a database using a set of keywords. Such structural information to be returned can be either trees or subgraphs representing how the objects, that contain the required keywords, are interconnected in a relational database or in an XML database. The structural keyword search is completely different from finding documents that contain all the user-given keywords. The former focuses on the interconnected object structures, whereas the latter focuses on the object content.

The book is organized as follows. In Chapter 1, we highlight the main research issues on the structural keyword search in different contexts. In Chapter 2, we focus on supporting structural keyword search in a relational database management system using the SQL query language. We concentrate on how to generate a set of SQL queries that can find all the structural information among records in a relational database completely, and how to evaluate the generated set of SQL queries efficiently. In Chapter 3, we discuss graph algorithms for structural keyword search by treating an entire relational database as a large data graph. In Chapter 4, we discuss structural keyword search in a large tree-structured XML database. In Chapter 5, we highlight several interesting research issues regarding keyword search on databases.

The book can be used as either an extended survey for people who are interested in the structural keyword search or a reference book for a postgraduate course on the related topics.

KEYWORDS

keyword search, interconnected object structures, relational databases, XML databases, data stream, rank

To my wife, Hannah
my children, Michael and Stephen

Jeffrey Xu Yu

To my wife, Michelle
my parents, Hanmin and Yaping

Lu Qin

To my parents, Qiyuan and Yumei

Lijun Chang

Contents

Preface

It has become highly desirable to provide flexible ways for users to query/search information by integrating database (DB) and information retrieval (IR) techniques in the same platform. On one hand, the sophisticated DB facilities provided by a database management system assist users to query well-structured information using a query language based on database schemas. Such systems include conventional RDBMSs (such as *DB2*, *ORACLE*, SQL-Server), which use SQL to query relational databases (*RDB*s) and *XML* data management systems, which use XQuery to query *XML* databases. On the other hand, IR techniques allow users to search unstructured information using keywords based on scoring and ranking, and they do not need users to understand any database schemas. The main research issues on DB/IR integration are discussed by Chaudhuri et al. [2005] and debated in a SIGMOD panel discussion [Amer-Yahia et al., 2005]. Several tutorials are also given on keyword search over *RDB*s and *XML* databases, including those by Amer-Yahia and Shanmugasundaram [2005]; Chaudhuri and Das [2009]; Chen et al. [2009].

The main purpose of this book is to survey the recent developments on keyword search over databases that focuses on finding *s*tructural information among objects in a database using a keyword query that is a set of keywords. Such structural information to be returned can be either trees or subgraphs representing how the objects, which contain the required keywords, are interconnected in an *RDB* or in an *XML* database. In this book, we call this *s*tructural keyword search or, simply, *k*eyword search. The structural keyword search is completely different from finding documents that contain all the user-given keywords. The former focuses on the interconnected object structures, whereas the latter focuses on the object content. In a DB/IR context, for this book, we use keyword search and keyword query interchangeably. We introduce forms of answers, scoring/ranking functions, and approaches to process keyword queries.

The book is organized as follows.

In Chapter 1, we highlight the main research issues on the structural keyword search in different contexts.

In Chapter 2, we focus on supporting keyword search in an RDMS using SQL. Since this implies making use of the database schema information to issue SQL queries in order to find structural information for a keyword query, it is generally called a schema-based approach. We concentrate on the two main steps in the schema-based approach, namely, how to generate a set of SQL queries that can find all the structural information among tuples in an *RDB* completely and how to evaluate the generated set of SQL queries efficiently. We will address how to find all or top-*k* answers in a static *RDB* or a dynamic data stream environment.

In Chapter 3, we also focus on supporting keyword search in an RDBMS. Unlike the approaches discussed in Chapter 2 using SQL, we discuss the approaches that are based on graph algorithms by materializing an entire database as a large data graph. This type of approach is called schema-free, in the sense that it does not request any database schema assistance. We introduce several algorithms, namely polynomial delay based algorithms, dynamic programming based algorithms, and Dijkstra shortest path based algorithms. We discuss how to find exact top-k and approximate top-k answers in a large data graph for a keyword query. We will discuss the indexing mechanisms and the ways to handle a large graph on disk.

In Chapter 4, we discuss keyword search in an *XML* database where an *XML* database is a large data tree. The two main issues are how to find all subtrees that contain all the user-given keywords and how to identify the meaning of such returned subtrees. We will discuss several algorithms to find subtrees based on lowest common ancestor (LCA) semantics, smallest LCA semantics, exclusive LCA semantics, etc.

In Chapter 5, we highlight several interesting research issues regarding keyword search on databases. The topics include how to select a database among many possible databases to answer a keyword query, how to support keyword query in a spatial database, how to rank objects according to their relevance to a keyword query using PageRank-like approaches, how to process keyword queries in an OLAP (On-Line Analytical Processing) context, how to find frequent additional keywords that are most related to a keyword query, how to interpret a keyword query by showing top-k SQL queries, and how to project a small database that only contains objects related to a keyword query.

The book surveys the recent developments on the structural keyword search. The book can be used as either an extended survey for people who are interested in the structural keyword search or a reference book for a postgraduate course on the related topics.

We acknowledge the support of our research on keyword search by the grant of the Research Grants Council of the Hong Kong SAR, China, No. 419109.

We are greatly indebted to M. Tamer Özsu who encouraged us to write this book and provided many valuable comments to improve the quality of the book.

Jeffrey Xu Yu, Lu Qin, and Lijun Chang
The Department of Systems Engineering and Engineering Management
The Faculty of Engineering
The Chinese University of Hong Kong
December, 2009