Information Sciences 288 (2014) 135–152

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A low redundancy strategy for keyword search in structured and semi-structured data

Jaime I. Lopez-Veyna *, Victor J. Sosa-Sosa, Ivan Lopez-Arevalo

Information Technology Laboratory, Center of Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV), Tamaulipas, Mexico

ARTICLE INFO

Article history:
Received 21 August 2013
Received in revised form 26 May 2014
Accepted 21 July 2014
Available online 7 August 2014

Keywords:
Keyword search
Virtual Document
Indexing
Database
Top-*k*

ABSTRACT

Keyword Search has been recognised as a viable alternative for information search in semi-structured and structured data sources. Current state-of-the-art keyword-search techniques over relational databases do not take advantage of correlative meta-information included in structured and semi-structured data sources leaving relevant answers out. These techniques are also limited due to scalability, performance and precision issues that are evident when they are implemented on large datasets. Based on an in-depth analysis of issues related to indexing and ranking semi-structured and structured information. We propose a new keyword-search algorithm that takes into account the semantic information extracted from the schemes of the structured and semi-structured data sources and combine it with the textual relevance obtained by a common text retrieval approach. The algorithm is implemented in a keyword-based search engine called *KESOSASD (Keyword Search Over Semi-structured and Structured Data)*, improving its precision and response time. Our approach models the semi-structured and structured information as graphs, and make use of a *Virtual Document Structure Aware Inverted Index (VDSAII)*. This index is created from a set of logical structures called *Virtual Documents*, which capture and exploit the implicit structural relationships (semantics) depicted in the schemas of the structured and semi-structured data sources. Extensive experiments were conducted to demonstrate that KESOSASD outperforms existing approaches in terms of search efficiency and accuracy. Moreover, KESOSASD is prepared to scale out and manage large databases without degrading its effectiveness.

1. Introduction

Thanks to the very intuitive and easy-to-use tools to publish information on the Web, nowadays Web search engines face a bigger challenge to organise and find relevant information that comes from different data sources. Additional to static web pages, in the Web, we can find a significant amount of information that has been stored in relational databases, XML documents or other data sources for along time. It is crucial for users to be able to search in a simple manner in these data sources.

Typing keywords as an input in a search engine has become the most used method to find information on the Web. The simplicity of search interfaces, the fact that users need neither any knowledge about database interrogation languages (e.g. SQL or XQuery) nor information that describes the schemes of the data sources to be queried are the main reasons for the

* Corresponding author. Tel.: +52 834 1070220; fax: +52 834 1070241.
E-mail addresses: jlopez@tamps.cinvestav.mx (J.I. Lopez-Veyna), vjsosa@tamps.cinvestav.mx (V.J. Sosa-Sosa), ilopez@tamps.cinvestav.mx (I. Lopez-Arevalo).

http://dx.doi.org/10.1016/j.ins.2014.07.054
0020-0255/

acceptance and popularity of keywords-based search. This simple search method has also become a viable alternative for information search in semi-structured and structured data sources like XML databases and relational databases respectively.

The research community has recognised the benefits of keyword search and has been working during the last ten years on projects to perform keyword search into relational databases [1–9], XML databases [10–16], graph databases, [17–20], and heterogeneous data sources [21–23]. The use of data graphs to model the scheme and the relations between tables or tuples in a databases becomes an alternative to perform keyword based queries over structured and semi-structured information. In this model the nodes represent the tuples (relational entities) and the edges represent the relationships between pairs of tuples (primary–foreign-key relationships). Nowadays it is getting most common to find applications that need to integrate structured or semi-structured data and text documents. One advantage when modelling data sources as data graphs is that the data graph keeps the values in redundant schema elements, and the search system does not need to access the underlying database once the data graph is constructed [6].

The state-of-the-art approaches can be classified into four main methods: (1) *Steiner Tree*, with proposals such as BLINKS [12], BANKS [2], EASE [23], DPBF [5], and CSTREE [8]. (2) *Candidate Networks*, with projects such as DISCOVER [4], DISCOVER II [24], XKeyword [10], SPARK [3] and SPARK II [25]. (3) *Tuple Units* that include EKSO [26] RETUNE [27] and SAINT [7]. And more recently (4) *Virtual Documents*, with proposals such as EKSO [26], SPARK [3] and KESOSD [9]. However these methods have some limitations. The Steiner Trees method is considered a NP-hard problem. Moreover, real databases can produce a large number of Steiner Trees, which are difficult to identify and index. The Candidate Network approach first needs to generate the Candidate Networks and then to evaluate them to find the best answer. The problem is that for a keyword query the number of Candidate Networks can be very large, and to find a common join expression to evaluate all the candidate networks could require a big computational effort. The use of Tuple Units in a general conception produce very large structures that most of the time store redundant information. Finally, the Virtual Documents approach produce in some cases redundant information in the Virtual Documents generated.

Inspired by the great success of the Information Retrieval approach on web search, other orthogonal proposals related of our work have emerged such as [28–32]. It is important to notice that our proposal represents an important improvement and evolution of the strategy used in [9].

1.1. Problem

Although keyword search has been proven to be effective for finding relevant documents, it presents some limitations on structured and semi-structured data that are not easy to carry out. The reason of this situation is that current Information Retrieval (IR) techniques applied on search engines have not been originally designed for this type of data sources [33]. They have commonly employed the inverted index to process keyword queries, which is effective for unstructured data but it is inefficient for semi-structured and structured data [21]. To answer a keyword-based query in relational databases, we need to include some slices of data that are often split across different tables. In a situation like this, if the tuples of these tables could be joined through their foreign–primary keys, it means that the information stored in these tuples are related. The tuples can belong to different tables and must be linked through primary or foreign key forming a unit of information. This unit of information can be indexed and used as a good method to find information using a keyword-based approach. However, with large databases, indexing units of information present an important issue because they can include a lot of redundant information that could make the search inefficient or intractable. Therefore, it is convenient to think about the right form to divide these units of information into parts that could be more compact, reducing the duplicated information.

The splitting of the information has the advantage of managing little slices of information that can be indexed in a more efficient way. We can consider these slices of information like documents that only contain the attributes with textual information. These documents contains the textual information of one or more tuples. In a related work [9], the author proposed an approach to identify and eliminate the redundancy generated in the documents created from tuples, producing efficient and accurate results, However, this strategy demands high storage consumption when processing large datasets due to excessive redundancy. In this context our proposal *Keyword Search Over Semi-structured and Structured Data (KESOSASD)* represents a scalable searching and indexing architecture that improves the technique applied in [9] with less storage requirements. The construction of these documents without redundancy is one of the motivation of this work. The structure of the information, extracted from the data sources, can be exploited and used for answering keyword queries.

Next-generation of keyword-based search for structured and semi-structured data requires the capability of integrating correlative meta-information, which represents some types of links between the information components and the relationships between the keywords. This meta-information can be useful for determining the relevance and the association of that information.

This paper proposes a new approach for *Keyword Search Over Semi-structured and Structured Data* called KESOSASD. KESOSASD makes an adequate query processing by using a classical IR metrics, but taking into account the semantic information extracted from the schemas provided by structured and semi-structured data sources. KESOSASD identifies the most relevant and meaningful tuples to answer keyword queries eliminating the redundancy in the Virtual Documents generated as answers. In our approach, we model semi-structured and structured data as graphs, where nodes can be tuples, and edges can be primary–foreign-key relationships. We enable adequate keyword search over structured and semi-structured data sources by the use of a *Virtual Document Structure Aware Inverted Index* called *VDSAII*. It uses an IR style that captures the structural relationships of the data that will be represented.