Comprehensive Study on Keyword Search on Semi Structured Data

Dayananda P., JSSATE, Bangalore, India
Sowmyarani C. N., RVCE, Bangalore, India

ABSTRACT

Keyword search is a user-friendly approach that enables inexperienced users to easily retrieve information from XML data with no specific knowledge of complex structured query language. Since an XML document can have a large size and contain a lot of information, an XML keyword search result should be a fragment of an XML document dynamically constructed at query time, which is achievable due to the structuredness of XML. Processing keyword searches on XML has several challenges, e.g., what are the elements in the XML document that are relevant to the query? How to generate the results efficiently and rank the results meaningfully? How to present the results to the user in a way such that the user can quickly find the desired information? In this survey, the authors review the papers in the literature that attempted to address these problems. The authors divide the existing approaches into several classes based on the problem they tackled, and perform a comprehensive analysis of these works.

KEYWORDS
Keyword Search, LCA Semantics, Lowest Common Ancestor, Semi Structured Data, XML

1. INTRODUCTION

With the development of information science and information society, an increasing number of information on the internet is stored in XML format. Given the fact that XML is becoming the standard in exchanging and representing data, how to extract accurate information in an effective and efficient way from XML documents has become an important issue arousing much interest of researchers. For any research oriented novel effort, the analysis of existing systems and respective in depth knowledge exploration is of great significance. The number of research methods are implemented in the field of information retrieval for keyword search on semi-structured data and these researches possesses certain strengths as well as limitation. Hence, understating the existing research for its respective strengths as well as weaknesses, the scholar can get better insight of the technologies and approaches to achieve optimal solution. Thus, taking into consideration of these requirements, in this chapter of the presented thesis, a number of literatures have been explored and studied for keyword search on semi-structured data.

1.1. Tree-based XML Keyword Search

When XML documents do not contain IDREF, they can be modeled as trees. Approaches to handle such documents are called tree-based approaches because they are based on tree model. Inspired by the hierarchical structure of the tree model, most of existing tree-based approaches are based on the LCA (Lowest Common Ancestor) semantics, which returns the lowest common ancestors of matching nodes to keyword queries.
There are many subsequent semantics to filter less meaningful answers. Existing works either improve the effectiveness by proposing a new semantics or improve the efficiency by proposing a new method for certain semantics. The widely accepted LCA-based semantics include LCA itself, SLCA, VLCA, MLCA, ELCA, and etc., among which, SLCA and ELCA are the most popular semantics. We classify the existing research works into these semantics and result of our classification is shown in Figure 1. Some research work involves study of more than one semantic such as XRANK (Guo, 2003), Set-intersection (Zhou, 2012), and Top-K (Guo, 2003).

1.1.1. LCA Semantics

The LCA semantics for XML keyword search was first proposed in XRANK (Guo, Shao, Botev et al., 2003, pp. 16-27). By the LCA semantics, for a set of matching nodes, each of which contains at least one query keyword and each query keyword matches at least one node in this set, the lowest common ancestor (LCA) of this set is a returned node. An answer is a subtree rooted as a returned node (i.e., an LCA) or a path from the returned node to matching nodes. XRANK is extended from Google’s Pagerank algorithm for ranking. It takes into account the proximity of the keywords and the references between attributes. XRANK implements a naive approach, and three optimized approaches afterwards to improve the search.

1.1.2. SLCA Semantics

The SLCA (Smallest LCA) semantics was first proposed in XKSearch (Xu, 2005). The SLCA semantics defines an SLCA to be an LCA that does not have any other LCAs as its descendants. There are many works on finding the set of SLCAs for a keyword query. XKSearch (Xu, 2005) proposes two efficient algorithms to compute SLCAs, namely Indexed Lookup Eager and Scan Eager. To find all SCLAs, there are two tasks, namely finding all LCAs and remove all ancestors among LCAs to get the SLCAs. It is costly to find all LCAs.

XKSearch optimizes as follows. Firstly, for each matching node u of the keyword which has the least number of matching nodes XKSearch finds its left and right match. The left (right) match v of u refers the matching node of the other queries or the keyword and among all nodes in u’s left (right) side, v is the nearest one (by pre-order). Only the LCA of u and v is a candidate SLCA. Thereby, it greatly reduces the number of computation of LCAs. The predominant characteristics of the approach SLCA refers that with provided two queries or keywords k1, k2 with one node u which comprises keyword k1, the one keyword seeks not to explore the complete keyword list so as to find the optimal solutions. In addition, one merely requires finding the left and right match of u in the k2 list, where the right or left match is the node with the minimal or maximum Dewey ID that is greater or smaller as compared to the Dewey ID of the node u. Multiway-SLCA (Sun, 2007) in addition attempted to enhance the performance of XKSearch scheme. The principal impetus behind this paradigm is the provision to avoid the redundancy of existing XKSearch by means of estimating the SLCAs through multiple intermediate SLCAs. Multi model SLCAs scheme estimates the individual SLCA by taking

Figure 1. Structural relationships among nodes

![Diagram](image-url)
Redesign of Home Care Service Delivery: A Systemic Approach to IT Innovations
www.igi-global.com/article/redesign-of-home-care-service-delivery/176941?camid=4v1a