Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems ●●● (●●●●) ●●●–●●●

FUZZY sets and systems

www.elsevier.com/locate/fss

Efficient keyword search in fuzzy XML

Jian Liu [a], X.X. Zhang [b]

[a] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China*
[b] *Northeastern University, Shenyang, 110004, China*

Received 9 March 2014; received in revised form 22 May 2016; accepted 23 May 2016

Abstract

Evaluation of keyword queries over XML documents is one of the most fundamental tasks for XML data management. Previous methods have focused on the processing of deterministic XML data. However, uncertain data are inherent in practical applications, and how to support efficient keyword search over fuzzy XML data remains at large an open problem. In this paper, we tackle the problem of efficiently producing SLCA (smallest lowest common ancestor) results for keyword queries in fuzzy XML documents. We propose an efficient approach that can find all SLCA results for a given keyword query over fuzzy XML data. In particular, we introduce an effective method to transform a simple keyword query into a segmented keyword query that captures the original query requirements and conforms to the underlying fuzzy XML data. The proposed approach could help us eliminate irrelevant SLCA results and speed up the query processing. The final experiments show the effectiveness and efficiency of our proposed approach in generating SLCA results.

Keywords: Keyword queries; Fuzzy XML; SLCA

1. Introduction

Over the past few years, the Extensible Markup Language (XML) has become the dominant data format for information exchange. With the proliferation of XML data comes the motivation to manipulate XML documents [1]. Traditionally, a structured query language, such as XPath [2] or XQuery [3], is used to search and manipulate XML data, which can convey complex semantic meanings and precisely retrieve the desired results [4]. For example, the BaseX [5] is an XML database system with XPath/XQuery processor. To issue a query which finds the XML sub-trees containing “federal republic inflation >10”, an XPath expression: intersect(//*[contains(.,“federal republic”)/ancestor::*,//inflation[.>10]/ancestor::*) can be used. Obviously, this XPath (or XQuery) expression requires users to know the complex XPath (or XQuery) query syntax. Nevertheless, due to the complexity of the query

E-mail address: lyot@163.com (J. Liu).

http://dx.doi.org/10.1016/j.fss.2016.05.015
0165-0114/© 2016 Elsevier B.V. All rights reserved.

syntax, it is difficult for a naïve user to specify an XPath (or XQuery) expression to express his or her search request [6,7].

Another alternative paradigm for searching XML data is the keyword search. As illustrated in [8], one of the key advantages of keyword search is its simplicity—users do not have to learn a complex query language. In other words, if a user is unfamiliar with the query syntax of XPath or XQuery, then a keyword search could be chosen for accessing XML data [9]. For example, to issue the example query above, a naïve user could input keywords "federal republic inflation >10" only, then the corresponding sub-tree answers (e.g., lowest common ancestor (LCA)) of these keywords could be automatically generated by using a keyword searching engine. Keyword search has emerged as a popular paradigm for information retrieval over XML data [4,6,8,10,11]. In [11], Xu and Papakonstantinou firstly gave the smallest lowest common ancestor (SLCA) semantics in XML keyword queries, and then introduced two algorithms for SLCA keyword queries over XML data. In [6], Li et al. proposed an estimation-based approach to compute the promising results for a keyword query over XML data sources, which can help a user quickly narrow down to his specific information need. In [4], Liu and Chen implemented a system that allowed users to search information in XML documents by keywords and identified meaningful returned nodes without the user solicitation. One of the significant merits of XML keyword queries is its simplicity—it allows users to find the information they are interested in without having to learn a complex query language. However, this kind of simple query technique may not be precise and can potentially return a large number of results that are not interesting to users. To address this problem, several refining attempts [12–14] incorporating some structural constraints are proposed. In [12], Lee et al. proposed a keyword search method that removed spurious results by using structural consistency. In [13], Petkova developed a query refinement technique that generated content-and-structure queries from plain-text queries. In [14], Termehchy and Winslett presented a ranking approach for keyword queries which avoided overreliance on shallow structural details.

In modern web applications that involve analysis and management of databases, uncertainty is often an inherent property of the data [15,16]. Traditional databases allow for the storage and retrieval of large amounts of data, but do not make any concessions for imprecision and uncertainty in the data [17]. In many domains, such as environmental surveillance, market analysis and quantitative economics research, it is difficult to state all information with one hundred percent certainty. This drives database researchers to develop specific solutions to provide supports for uncertain XML data processing. As a result, researches on uncertain XML data management are extensively under way. Some representations of probabilistic data in XML were proposed in previous works, such as Nierman et al. [18] and Hung et al. [19]. Unfortunately, although fuzzy values have been employed to model and handle imprecise information in relational databases since Zadeh introduced the theory of fuzzy sets [20], only relatively little work has been carried out in extending and querying XML towards the representation of the fuzzy set concepts.

As illustrated in [22], a fuzzy XML document (fuzzy XML for short) is an XML document that contains possibility information quantified by using fuzzy sets. Its theoretical foundation is possibility theory,[1] which is an uncertainty theory devoted to the handling of incomplete information. In fact, fuzzy XML has a long history based on the notion of fuzzy trees [21], comparable to the one of the fuzzy relational model. In real-world scenarios, much of the data in the types of applications where uncertainty is an issue, such as web data and scientific data, are not easy to represent in a relational model, even ignoring issues of uncertainty [18]. The arising of fuzzy XML is driven by application needs that involve data not readily amenable to a relational representation, and all of the advantages of fuzzy XML naturally result in the boom in the research of fuzzy XML databases [22].

Invariably, manipulating fuzzy XML documents to retrieve useful information is always an obstacle. Previous studies mainly focused on twig queries [23–26], with little light shed on keyword queries over fuzzy XML data. The major problem of using twig queries is that, in order to return the desired query results, a user has to know the schema of the fuzzy XML documents and submit a complex structural query. If the schema of the data is unavailable, complicated or fast evolving, then a keyword search, which needs users to submit a few keywords and then a database management system will automatically find some suitable fragments from the fuzzy XML data, could be chosen for searching XML data. Supporting keyword search over fuzzy XML data is significant and promising, because it enables inexperienced users to easily search fuzzy XML database with no specific knowledge of complex structural query languages. In keyword search over uncertain XML data, the focus is not merely on finding relevant fragments

[1] http://www.scholarpedia.org/article/Possibility_theory.

Please cite this article in press as: J. Liu, X.X. Zhang, Efficient keyword search in fuzzy XML, Fuzzy Sets Syst. (2016), http://dx.doi.org/10.1016/j.fss.2016.05.015