International Journal of Advance Research in Science and Engineering
Vol. No.5, Special Issue No. 01, March 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

SURVEY PAPER ON XML CONTEXT DIVERSIFIED SEARCH

Simon Farande[1], Dr.D.S.Bhosale[2]

[1]Student, [2]Guide, Ashokrao Mane Group of Institutions, Vathar, Maharashtra, (India)

ABSTRACT

When user searches a keyword he types short and vague keywords and the ambiguity keyword query makes it difficult to effectively answer keyword queries. This paper focuses on diversified XML keyword search based on its different contexts in the XML data. If the user types short and vague query it is searched in XML data, and derive keyword search candidates of the query by a simple feature selection model. Then design an effective XML keyword search diversification model by implementing two algorithms.

Keywords: Context, Diversified, Query, Search, Xml

I. INTRODUCTION

[1]Keyword search on structured and semi-structured data has attracted much research interest recently, as it enables users to retrieve information without the need to learn sophisticated query languages and database structure . Compared with keyword search methods in Information Retrieval (IR) that prefer to find a list of relevant documents, keyword search approaches in structured and semi-structured data (denoted as DB&IR) concentrate more on specific information contents, e.g., fragments rooted at the smallest lowest common ancestor (SLCA) nodes of a given keyword query in XML. Given a keyword query, a node v is regarded as an SLCA if(1) the subtree rooted at the node v contains all the keywords, and(2) there does not exist a descendant node v′ of v such that the subtree rooted at v′ contains all the keywords.

In other words, if a node is an SLCA, then its ancestors will be definitely excluded from being SLCAs, by which the minimal information content with SLCA semantics can be used to represent the specific results in XML keyword search. In this system the well accepted SLCA semantics as a result metric of keyword query over XML data is to be adopted. In general, the more keywords a user's query contains, the easier the user's search intention with regards to the query can be identified. However, when the given keyword query only contains a small number of vague keywords, it would become a very challenging problem to derive the user's search intention due to the high ambiguity of this type of keyword queries. Although sometimes user involvement is helpful to identify search intentions of keyword queries, a user's interactive process may be time consuming when the size of relevant result set is large. To address this, develop a method of providing diverse keyword query suggestions to users based on the context of the given keywords in the data to be searched. By doing this, users may choose their preferred queries or modify their original queries based on the returned diverse query suggestions.

When the given keyword query only contains a small number of vague keywords, it would become a very challenging problem to derive the user's search intention due to the high ambiguity of this type of keyword queries. Although sometimes user involvement is helpful to identify search intentions of keyword queries, a user's interactive process may be time consuming when the size of relevant result set is large. To address this, **Jianxin Li, Chengfei Liu *Member,* and Jeffrey Xu Yu** [1] proposed a method of providing diverse keyword query suggestions to users based on the context of the given keywords in the data to be searched. By doing this, users may choose their preferred queries or modify their original queriesbased on the returned diverse query suggestions.

The most straightforward and effective querying method for non-structured document collections is the well-known keyword search. One of its key advantages is simplicity, since users only need to specify the keywords they are interested in. However, XML document collections have both content and structure, and may be queried by content, structure or both.**Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, and Eduardo Vicente-López**[2], keep the simple keyword search query interface, although they exploit XML structure during the query processing, so that the retrieved results can be any kind of document components.

Data Model and classification of nodes can be done as proposed by **Youqiang Guo , Guixiu Tao, Yuqing Liang, Lei Wang, Honghao Zhu[3]**

XML documents, the subtrees connected by the content nodes that directly contain some keywords are not easily retrieved. Accordingly, it is much harder to retrieve the integrated subtrees (like the documents and interrelated tuples) connected by the content nodes and complementary nodes that do not contain any keyword but contain some relevant and meaningful data, since it is rather difficult to determine which nodes are complementary nodes. **LI Guoliang ,FENG Jianhua And ZHOU Lizhu**[4]mentioned how to retrieve those compact, integral subtrees, called self-integral trees (SI-Trees), which contain complementarynodes that capture the focus of keyword queries, besides the content nodes. More importantly, each self-integral tree represents an integrated meaning to answer a keyword query.

Jianxin Li, Chengfei Liu, Rui Zhou and Bo Ning[5] suggest Given a keyword query q and an XML schema tree T, a set of structured queries Q may be constructed and evaluated over the data source conforming to T for answering q. The answer to the XML keyword query q may be a big number of relevant XML fragments. In contrast, the answer to the top-k keyword query is an ordered set of fragments, where the ordering reflects how closely each fragment matches the given keyword query. Therefore, only the top k results with the highest relevance w.r.t. q need to be returned to users.

II.FEATURE SELECTION AND DIVERSIFICATION

[1]When the document is uploaded to the search engine its information is stored in XML file (T) and the relevance based term pair dictionary (W) is also stored in that file. At the time of search the distinct term-pairs are selected based on their mutual information .Mutual information has been used as a criterion for feature selection and feature transformation in machine learning. It can be used to characterize both the relevance and redundancy of variables, such as the minimum redundancy feature selection.

Mutual information (MI) is calculated by formula

MI (x,y,T)=Prob(x,y,T)*log [Prob(x,y,T)/(Prob(x,T)*Prob(y,T))]

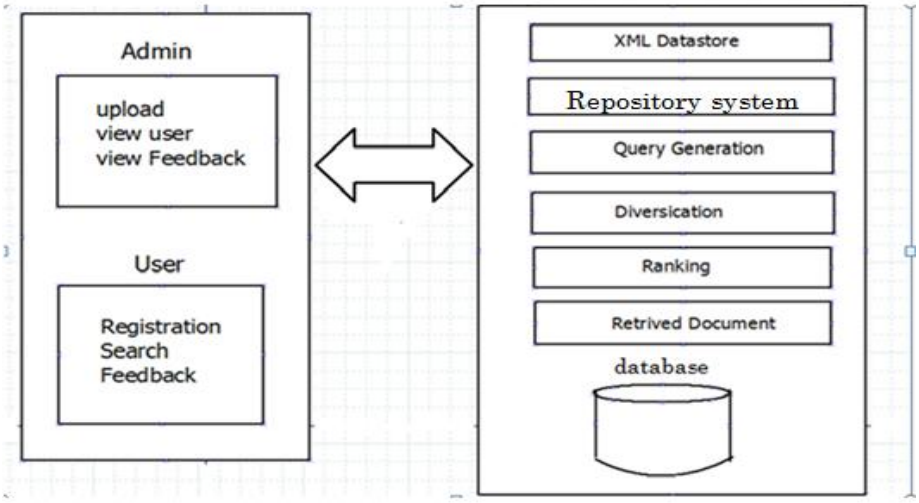

Fig.1 Architecture of XML context diversified search

a. Baseline Solution

[1]Given a keyword query, the intuitive idea of the baseline algorithm is to first retrieve the relevant feature terms with high mutual scores from the term correlated graph of the XML data T ; then generate list of query candidates that are sorted in the descending order of total mutual scores; and finally compute the SLCAs as keyword search results for each query candidate and measure its diversification score. As such, the top-k diversified query candidates and their corresponding results can be chosen and returned.

b. Anchor-based Pruning Solution

[1]To reduce the computational cost, design an anchor-based pruning solution, which can avoid the unnecessary computational cost of unqualified SLCA results (i.e., duplicates and ancestors). First analyze the interrelationships between the intermediate SLCA candidates that have been already computed for the generated query candidates Q and the nodes that will be merged for answering the newly generated query candidate qnew. And then, propose the detailed description and algorithm of the anchor-based pruning solution.

III. METHODOLOGY

[1]A real dataset, DBLP and a synthetic XML benchmark dataset XMark can be used for testing the proposed XML keyword search diversification model and designed algorithms. Compared with DBLP dataset, the synthetic XMark dataset has varied depths and complex data structures, but it does not contain clear semantic information due to its synthetic data.Therefore, only use DBLP dataset to measure the effectiveness of diversification model in this work. For each XML dataset used, select some terms based on the following criteria: (1) a selected term should often

appear in user-typed keyword queries;(2) a selected term should highlight different semantics when it co-occurs with feature terms in different contexts.

REFERENCES

[1] Jianxin Li, Chengfei Liu *and Jeffrey Xu Yu* "Context-based Diversification for Keyword Queries over XML Data" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING , VOL. 26, NO. 5, MAY 2014

[2] Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, and Eduardo Vicente-López "Using Personalization to Improve XML Retrieval" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 5, MAY 2014

[3] Youqiang Guo , Guixiu Tao, Yuqing Liang, Lei Wang, Honghao Zhu, " XML Keyword Search Based on Node Classification and Hierarchical Semantics" Communications in Information Science and Management Engineering Jan. 2014, Vol. 4 Iss. 1, PP. 6-12

[4] LI Guoliang ,FENG Jianhua And ZHOU Lizhu," Keyword Searches in Data-Centric XML Documents Using Tree Partitioning" in TSINGHUA SCIENCE AND TECHNOLOGY ISSN 1007-0214 02/ 21 pp7-18 Volume 14, Number 1, February 2009

[5] Jianxin Li, Chengfei Liu, Rui Zhou and Bo Ning ," Processing XML Keyword Search by Constructing Effective Structured Queries" Advances in Data and Web Management Lecture Notes in Computer Science Volume 5446 , 2009, pp 88-9