International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 14
October 2016

Diversify Based On the Context of the Word Queries for Xml Data

[1]Mrs.Joshi Padma Narasimhachari, [2]Kaveti Sai Teja ,[3]Dr. N. Ravi Shankar,[4]Dr. M. B. Raju

[1]Associate Professor,Head Of the Department,CSE, Sreyas Institute of Engineering & Technology, padmajoshi2015@gmail.com

[2]PG Scholar, Department of CSE, Sreyas Institute of Engineering & Technology, saiteja.kaveti@gmail.com

[3]Professor, Department of CSE, Lakireddy Balreddy College of Engineering, Vijayawada .ravish00@yahoo.com

[4]Professor, Department of CSE, KrishnaMurthy Institute of Engineering & Technology, Hyderabad. drrajucse@gmail.com

ABSTRACT

While keyword query empowers ordinary users to search vast amount of data, the ambiguity of keyword query makes it difficult to effectively answer keyword queries, especially for short and vague keyword queries. To address this challenging problem, in this paper we propose an approach that automatically diversifies XML keyword search based on its different contexts in the XML data. Given a short and vague keyword query and XML data to be searched, we first derive keyword search candidates of the query by a simple feature selection model. And then, we design an effective XML keyword search diversification model to measure the quality of each candidate. After that, two efficient algorithms are proposed to incrementally compute top-k qualified query candidates as the diversified search intentions. Two selection criteria are targeted: the k selected query candidates are most relevant to the given query while they have to cover maximal number of distinct results. At last, a comprehensive evaluation on real and synthetic data sets demonstrates the effectiveness of our proposed diversification model and the efficiency of our algorithms.

1.INTRODUCTION

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 [1]. Compared with keyword search methods in information retrieval (IR) that prefer to find a list of relevant documents, keyword search approaches in structured and semistructured data (denoted as DB and 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 v0 of v such that the subtree rooted at v0 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 paper, we adopt the well-accepted LCA semantics as a result metric of keyword query over XML data.

2.PROBLEM DEFINITION

Given a keyword query q and an XML data T, our target is to derive top-k expanded query candidates in terms of high relevance and maximal diversification for q in T. Here, each query candidate represents a context or a search intention of q in T.

2.1 Feature Selection Model

Consider an XML data T and its relevance-based term-pair dictionary W. The composition method of W depends onthe application context and will not affect our subsequent discussion. As an example, it can simply be the full or a subset of the terms comprising the text in T or a well-specified set of term-pairs relevant to some applications.

In this work, the distinct term-pairs are selected based on their mutual information as [15], [16]. 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.

Assume we have an XML tree T and its sample result set be the probability of term x appearing

If terms x and y are independent, then knowing x does not give any information about y and vice versa, so their mutual information is zero. At the other extreme, if terms x and y are identical, then knowing x determines the value of y and vice versa. Therefore, the simple measure can be used to quantify how much the observed word co-occurrences maximize the dependency of feature terms while reduce the redundancy of feature terms. In this work, we use the popularly-accepted mutual information model as follows:

for the query keywords in q. Each combination of the feature terms in matrix may represent a search intention with the specific semantics. For example, the

combination "query expansion database systems" targets to search the publications discussing the problem of query expansion in the area of database systems, e.g., one of the works, "query expansion for information retrieval" published in Encyclopedia of Database Systems in 2009, will be returned. If we replace the feature term "systems" with "relational", then the generated query will be changed to search specific publications of query expansion over relational database, in which the returned results are empty because no work is reported to the

3.EXISTING SYSTEM

The problem of diversifying keyword search is firstly studied in IR community. Most of them perform diversification as a post-processing or reranking step of document retrieval based on the analysis of result set and/or the query logs. In IR, keyword search diversification is designed at the topic or document level. Liu et al. is the first work to measure the difference of XML keyword search results by comparing their feature sets. However, the selection of feature set is limited to metadata in XML and it is also a method of post-process search result analysis.

4.PROPOSED SYSTEM

To address the existing issues, we will 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. To address the existing limitations and challenges, we initiate a formal study of the diversification problem in XML keyword search, which can directly compute the diversified results without retrieving all the relevant candidates. Towards this goal, given a keyword query, we first derive the co-related feature terms for each query keyword from XML data based on mutual information in the probability theory, which has been used as a criterion for feature selection. The selection of our feature terms is not limited to the labels of XML elements.

5.EXTRACTING FEATURE TERMS

To address the problem of extracting meaningful feature terms w.r.t. an original keyword query, there are two relevant works [17], [18]. In [17], Sarkas et al. proposed a solution of producing top-k interesting and meaningful expansions to a keyword query by extracting k additional words with high "interestingness" values. The expanded queries can be used to search more specific documents. The interestingness is

formalized with the notion of surprise [19], [20], [21]. In [18], Bansal et al. proposed efficient algorithms to identify keyword clusters in large collections of blog posts for specific temporal intervals. Our work integrates both of their ideas together: we first measure the correlation of each pair of terms using our mutual information model in Equation (1), which is a simple surprise metric; and then we build term correlated graph that maintains all the terms and their correlation values. Different from [17], [18], our work utilizes entity-based sample information to build a correlated graph with high precision for XML data. In order to efficiently measure the correlation of a pair of terms, we use a statistic method to measure how much the co-occurrences of a pair of terms deviate from the independence assumption where the entity nodes (e.g., the nodes with the "*" node types in XML DTD) are taken as a sample space. For instance, given a pair of terms x and y, their mutual information score can be calculated based on Equation (1) where Probðx; TÞ (or Probðy; TÞ) is the value of dividing the number of entities containing x (or y) by the total entity size of the sample space; Probðfx; yg; TÞ is the value of dividing the number of entities containing both x and y by the total entity size of the sample space.

In this work, we build a term correlated graph offline, that is we precompute it before processing queries. The correlation values among terms are also recorded in the graph, which is used to generate the term-feature dictionary W. During the XML data tree traversal, we first extract the meaningful text information from the entity nodes in XML data. Here, we would like to filter out the stop words. And then we produce a set of term-pairs by scanning the extracted text. After that, all the generated term-pairs will be recorded in the term correlated graph. In the procedure of building correlation graph, we also record the count of each term-pair to be generated from different entity nodes. As such, after the XML data tree is traversed completely, we can compute the mutual information score for each termpair based on Equation (1). To reduce the size of correlation graph, the term-pairs with their correlation lower than a threshold can be filtered out. Based on the offline built graph, we can on-the-fly select the top-m distinct terms asits features for each given query keyword.

6.KEYWORD SEARCH DIVERSIFICATION ALGORITHMS

In this section, we first propose a baseline algorithm toretrieve the diversified keyword search results. And then,two anchor-based pruning algorithms are designed toimprove

the efficiency of the keyword search diversificationby utilizing the intermediate results.

6.1.Baseline Solution

Given a keyword query, the intuitive idea of the baselinealgorithm is to first retrieve the relevant feature terms withhigh mutual scores from the term correlated graph of the XML data T; then generate list of query candidates that aresorted 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. Different from traditional XML keyword search, our work needs to evaluate multiple intended query candidates and generate a whole result set, in which the results should be diversified and distinct from each other. Therefore, we have to detect and remove the duplicated or ancestor SLCA results that have been seen when we obtain new generated results. The detailed procedure has been shown in Algorithm 1.

Given a keyword query q with n keywords, we first load its pre-computed relevant feature terms from the term correlated graph G of XML data T, which is used to construct a matrix Mm_n as shown in line

1. And then, we generate a new query candidate qnew from the matrix Mm_n by calling the function GenerateNewQuery() as shown in line

2. The generation of new query candidates are in the descending order of their mutual information scores. lines 3-7 show the procedure of computing Probðq j qnew; TÞ. To compute the SLCA results of qnew, we need to retrieve the precomputed node lists of the keyword-feature term pairs in qnew from T by getNodeListðsixjy; TÞ. Based on the retrieved node lists, we can compute the likelihood of generating the observed query q while the intended query is actually qnew, After that, we can call for the function ComputeSLCA() that can be implemented using any existing XML keyword search method. In lines 8-16, we compare the SLCA results of the current query and the previous queries in order to obtain the distinct and diversified SLCA results. At line 17, we compute the final score of qnew as adiversified query candidate w.r.t. the previously generated query candidates in Q. At last, we compare the new query and the previously generated query candidates and replace the unqualified ones in Q, which is shown in lines 18-23.

After processing all the possible query candidates, we can return the top k

generated query candidates with their SLCA results.

Algorithm 1. Baseline Algorithm

input: a query q with n keywords, XML data T and its term correlated graph G output: Top-k search intentions Q and the whole result set F

1: Mm_n ¼ getFeatureTerms(q, G);

2: while (qnew ¼ GenerateNewQuery (Mm_n)) 6¼ null do

3: f ¼ null and prob s k ¼ 1;

4: lixjy ¼ getNodeList(sixjy , T) for sixjy 2 qnew ^ 1 _ ix _ m ^ 1 _ jy _ n;

5: prob s k ¼ Q fixjy2sixjy2qnew ð jlixjy j getNodeSizeðfixjy ;T ÞÞ;

6: f ¼ ComputeSLCA({lixjyg);

7: prob q new ¼ prob s k * jfj;

8: if F is empty then

9: scoreðqnewÞ ¼ prob q new;

10: else

11: for all Result candidates rx 2 f do

12: for all Result candidates ry 2 F do

13: if rx ¼¼ ry or rx is an ancestor of ry then

14: f:removeðrxÞ;

15: else if rx is a descendant of ry then

16: F:removeðryÞ;

17: scoreðqnewÞ ¼ prob q new * jfj* jfj

18: if jQj < k then

19: put qnew : scoreðqnewÞ into Q;

20: put qnew : f into F;

21: else if scoreðqnewÞ > scoreðfq0new 2 QgÞ then

22: replace q0new : scoreðq0newÞ with qnew : scoreðqnewÞ;

23: F:removeðq0newÞ;

24: return Q and result set F;

In the worst case, all the possibe queries in the matrix may have the possibility of being chosen as the top-k qualified query candidates. In this worst case, the complexity of the algorithm is O(mjqj _ L1 P jqj i¼2 logLi) where L1 is the shortest node list of any generated query, jqj is the number of original query keywords and m is the size of selected features for each query keyword. In practice, the complexity of the algorithm can be reduced by reducing the number m of feature terms, which can be used to bound the number (i.e., reducing the value of mjqj) of generated query candidates.

7.CONCLUSION

In this paper, we first presented an approach to search diversified results of keyword query from XML data based on the contexts of the query keywords in the data. The diversification of the contexts was measured by exploring their relevance to the original query and the novelty of their results. Furthermore, we designed three efficient algorithms based on the observed properties of XML keyword search results.

8 REFERENCE

[1] Y. Chen, W. Wang, Z. Liu, and X. Lin, "Keyword search on structured and semi-structured data," in Proc. SIGMOD Conf., 2009, pp. 1005–1010.

[2] L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram, "Xrank: Ranked keyword search over xml documents," in Proc. SIGMOD Conf., 2003, pp. 16–27.

[3] C. Sun, C. Y. Chan, and A. K. Goenka, "Multiway SLCA-based keyword search in xml data," in Proc. 16th Int. Conf. World Wide Web, 2007, pp. 1043–1052.

[4] Y. Xu and Y. Papakonstantinou, "Efficient keyword search for smallest lcas in xml databases," in Proc. SIGMOD Conf., 2005, pp. 537–538.

[5] J. Li, C. Liu, R. Zhou, and W. Wang, "Top-k keyword search over probabilistic xml data," in Proc. IEEE 27th Int. Conf. Data Eng., 2011, pp. 673–684.

[6] J. G. Carbonell and J. Goldstein, "The use of MMR, diversitybased reranking for reordering documents and producing summaries," in Proc. SIGIR, 1998, pp. 335–336.

[7] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," in Proc. 2nd ACM Int. Conf. Web Search Data Mining, 2009, pp. 5–14.

[8] H. Chen and D. R. Karger, "Less is more: Probabilistic models for retrieving fewer relevant documents," in Proc. SIGIR, 2006, pp. 429–436.

[9] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. B€uttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," in Proc. SIGIR, 2008, pp. 659–666.

[10] A. Angel and N. Koudas, "Efficient diversity-aware search," in Proc. SIGMOD Conf., 2011, pp. 781–792.

AUTHOR'S PROFILE:

Mrs.Joshi Padma Narasimhachari
Associate Professor,Head Of the Department,CSE, Sreyas Institute of Engineering & Technology,
padmajoshi2015@gmail.com

Kaveti Sai Teja
PG Scholar, Department of CSE, Sreyas Institute of Engineering & Technology, saiteja.kaveti@gmail.com