Context-Based Diversification for Keyword Queries over XML Data

Naresh. M

Department of CSE, Newtons Institute of Engineering, Macherala, JNTUK, AP, India

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.

I. INTRODUCTION

Keyword search on structured and semi-structured data has attracted much research interest recently, as it enables common users to retrieve information from such structured data sources without the need to learn sophisticated query languages and database structure. In general, the more keywords a given keyword query contains, the easier the search semantics of the keyword query can be identified. However, when the given keyword query only contains a small number of vague keywords, it will become a very challenging problem to derive the search semantics of the query due to the high ambiguity of this type of keyword queries. Although Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. sometimes user involvement is helpful to identify search semantics of keyword queries, it is not always applicable to rely on users be-cause the keyword queries may also come from system application. In this application case, web or database search engine may need to automatically compute the search semantics of short and frequent keyword queries only based on the data to be searched. The derived search semantics will be maintained and updated in an off-line way. Once a keyword query is issued by the real users, its corresponding search semantics can be directly used to make an instant response. In this paper, we mainly pay attention to the problem of effectively deriving the search semantics of keyword queries with the consideration of data only, which does not receive much closer attention in the previous works.

| Table 1 |
| Top 10 Selected Feature Terms of q |
|---|
| Database | system; relational; protein; distributed; oriented; image; sequence; search; model; large. |
| Query | language; expansion; optimization; evaluation; complexity; log; efficient; distributed; semantic; translation. |

II. 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.

A. Feature Selection Model.

Consider an XML data T and its relevance-based term-pair dictionary W. The composition method of W depends on the 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. Mutual information has been used as a criterion for feature selection and feature transformations in machine learning. It can be used to characterize both the relevan ce and redundancy of variables, such as the minimum redundancy feature selection. Assume we have an XML tree T and its sample result set
Let \( \text{Prob}(x, T) \) be the probability of term \( x \) appearing in \( R(T) \), i.e., \( \text{Prob}(x, T) = |R(x,T)|/|R(T)| \) where \( |R(x,T)| \) is the number of results containing \( x \). Let \( \text{Prob}(x, y, T) \) be the probability of terms \( x \) and \( y \) co-occurring in \( R(T) \), i.e., \( \text{Prob}(x, y, T) = |R(x,y,T)|/|R(T)| \). 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 by how much the observed word co-occurrences that 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.

\[
MI(x, y, T) = \text{Prob}(x, y, T) \times \log \frac{\text{Prob}(x, y, T)}{\text{Prob}(x, T) \times \text{Prob}(y, T)}
\]

### TABLE 2

| database | system | relational | protein | distributed | oriented |
|----------|--------|------------|---------|-------------|----------|
| 7.06     | 3.84   | 2.79       | 2.25    | 2.06        |

| Mutual score (log) |
|--------------------|
| image | sequence | search | small | large |
| 1.23 | 1.31 | 1.1 | 1.04 | 1.02 |

| Mutual score (log) |
|--------------------|
| language | expansion | optimization | evaluation complexity |
| 3.63 | 2.97 | 2.3 | 1.71 | 1.41 |

| Mutual score (log) |
|--------------------|
| log | efficient | distributed | semantic translation |
| 1.17 | 1.03 | 0.99 | 0.86 | 0.70 |

For each term in the XML data, we need to find a set of feature terms where the feature terms can be selected in any way, e.g., top \( m \) feature terms or the feature terms with their mutual values higher than a given value based on domain applications or data administrators. The feature terms can be pre-computed and stored before the procedure of query evaluation. Thus, given a keyword query, we can obtain a matrix of features for the query keywords against the XML data to be searched. The matrix constructs a space of search intentions of the original query w.r.t. the XML data. Therefore, our first problem is to find a set of feature terms from the matrix, which has the highest probability of interpreting the contexts of original query. In this work, we extract and evaluate the feature terms at the entity level of XML data.

### III. EXISTING SYSTEMS

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.

#### A. Disadvantage of existing system

- 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.
- It is not always easy to get this useful taxonomy and query logs. In addition, the diversified results in IR are often modeled at document levels.
- A large number of structured XML queries may be generated and evaluated.
- There is no guarantee that the structured queries to be evaluated can find matched results due to the structural constraints.
- The process of constructing structured queries has to rely on the metadata information in XML data.

### IV. 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.

Each combination of the feature terms and the original query keywords may represent one of diversified contexts (also denoted as specific search intentions). And then, we evaluate each derived search intention by measuring its relevance to the original keyword query and the novelty of its produced results.

To efficiently compute diversified keyword search, we propose one baseline algorithm and two improved
algorithms based on the observed properties of diversified keyword search results.

A. Advantages of proposed system.

✓ Reduce the computational cost.
✓ Efficiently compute the new SLCA results.
✓ We get that our proposed diversification algorithms can return qualified search intentions and results to users in a short time.

A. System Architecture.

1) Modules:

1. Admin
2. User
3. XML Query Answering

Admin:
Admin maintains the total information about the whole application. Admin maintains the data in XML format only.

User:
User searches queries and gets the reply in XML format.

XML Query Answering:
In this project, users search for information in semi-

![Fig.1. System Architecture](image)

V. KEYWORD SEARCH DIVERSIFICATION ALGORITHM

In this section, we first introduce the procedure of generating a new query from the matrix of the original keyword query w.r.t. the data to be searched. And then based on the matrix, we propose a baseline algorithm to retrieve the diversified keyword search results. At last, two anchor-based pruning algorithms are designed to improve the efficiency of the keyword search diversification by utilizing the intermediate results.

A. Generate Search Intentions.

Given a keyword query q, we first retrieve the corresponding feature terms for each query keyword and then construct a matrix of search intentions. In the matrix, the feature terms in each column are sorted based on their mutual information scores. Each combination of the feature terms (one term per column) represents a search intention. We iteratively choose the combination with the maximal aggregated mutual information score as the next best search.

B. Baseline Solution.

Given a keyword query, the intuitive idea of baseline algorithm is that we first retrieve the pre-computed feature terms of the given keyword query from the XML data T and then generate all the possible intended queries based on the retrieved feature terms; at last, we compute the SLCAs as keyword search results for each query and measure its diversification score. As such, the top-k diversified queries and their corresponding results can be returned to users.

Algorithm 1. Baseline Algorithm

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

1: M_{max} = getFeatureTerms(q, G);
2: while (q_{new} = GenerateNewQuery(M_{max}) ≠ null do
3:     φ = null and prob_{knew} = 1;
4:     l_{m,n} = getNodeList(s_{m,n}) for s_{m,n} ∈ q_{new} ∧ 1 ≤ l_{m,n} ≤ m ∧ 1 ≤ n ≤ n;
5:     prob_{knew} = \prod_{l_{m,n} \in q_{new} \wedge l_{m,n} \leq n} \text{hit}_{m,n}(G);
6:     φ = ComputeSLCA(l_{m,n});
7:     prob_{knew} = prob_{knew} \cdot \text{prob}_{knew}^*; φ;
8:     if φ is empty then
9:         score(q_{new}) = prob_{knew};
10:     else
11:         for all Result candidates r_{c} ∈ φ do
12:             if r_{c} = r_{c} or r_{c} is an ancestor of r_{c} then
13:                 φ.remove(r_{c});
14:         if r_{c} is a descendant of r_{c} then
15:             φ.remove(r_{c});
16:         if |Q| < k then
17:             score(q_{new}) = prob_{knew} \cdot \text{score}(q_{new})^*; φ;
18:         put q_{new} : score(q_{new}) into Q;
19:         put φ into Φ;
20:     else if score(q_{new}) > score(q_{new}^*) then
21:         replace q_{new} : score(q_{new}) with q_{new} : score(q_{new})^*;
22:         φ.remove(q_{new});
23:         φ.remove(q_{new});
24:         return Q and result set Φ;
```

VI. RELATED WORK

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. Each combination of the feature terms and the original query keywords may represent one of diversified contexts (also denoted as specific
search intentions). And then, we evaluate each derived search intention by measuring its relevance to the original keyword query and the novelty of its produced results. To efficiently compute diversified keyword search, we propose one baseline algorithm and two improved algorithms based on the observed properties of diversified keyword search results.

VII. 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. Finally, we verified the effectiveness of our diversification model by analyzing the returned search intentions for the given keyword queries over DBLP data set based on the nDCG measure and the possibility of diversified query suggestions. Meanwhile, we also demonstrated the efficiency of our proposed algorithms by running substantial number of queries over both DBLP and XMark data sets. From the experimental results, we get that our proposed diversification algorithms can return qualified search intentions and results to users in a short time.

REFERENCES

[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, diversity-based 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. Bütcher, 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.