Anaphora Method Based Context-Diversification for Keyword Queries

Mrudula.M.S[1], Anoop.S[2]

[1]*(Department of Computer Science, College of Engineering Perumon, Kollam, India)*

[2]*(Department of Information Technology, College of Engineering Perumon, Kollam, India)*

Abstract: *Keyword based searching is the important part of research domain. The search can be applied on structured and /or semi-structured information .Queries are used to fetch large amount of data from databases. Due to ambiguity problem system can't effectively answer short and vague queries. For solving this ambiguity problem different context in XML data is used. Here XML keyword diversification model with the help of Anaphora method is proposed .Anaphora is the phenomenon of referring to an antecedent (Metonymically also referring expression) Subtype are pronouns and definite NPs. Two selection criteria are targeted: the k selected query candidates are most pertinent to the given query while they have to cover maximal number of distinct results. Judgment on real and synthetic data sets bespeak the fruitfulness of our proposed diversification model and the performance of our algorithms.*

Keywords: *Anaphora method, XML Data, keyword query.*

I. Introduction

In our day today life the searching of information is become a part of our working. When we search for some information or data's search engine gives the high resultant data's. But this result contains the relevant and irrelevant document. Keyword search approach in structured and semi structured data concentrate more on information content. Keyword search for smallest lowest common ancestors (SLCAs) in XML data has recently been proposed as a meaningful way to identify affecting data nodes in XML data where their sub trees contain an input set of keywords. Here well authorized SLCA semantic is adopted.

It is easy to retrieve information when the query contain more keyword otherwise it is difficult task due to ambiguity of keyword query. Some time user interaction is helpful for search intension. Problem with user interaction is that it is time consuming to solve this problem develop a method of providing diverse keyword suggestion based on the context that we are searched.

The process here is that when we given a query first identifies the keywords from the query for getting appropriate result. Then for each keyword abstract the related feature terms from a given XML data set based on predefined metadata and its features. This process is similar to the feature selection. The selected feature terms is not same as the labels of XML elements. Each unique combination of the feature terms and query keywords may enact one of diversified contexts. After scrutinizing the context of diversified query in terms of its relevance with actual query and novelty of produced result we will get proper queries. When we work with large xml data T, our aim is to derive top-k expanded query candidates from a given query Q which has more important and maximal Heterogeneity where every candidate in candidate list represent the search intention of q in T. To effectively calculate results a method here used is called anaphora that is in case of multiple query by using Contextual meaning of one another can optimize. Two algorithms are proposed baseline algorithm and two anchor based pruning algorithm.

II. Existing System

This for structured and semi-structured data, various techniques are used for keyword search. Query optimization , ranking phases , top crucial query processing is discussed. Different data models such as XML , graph-structured data is discussed. Application of these concepts are also figured in which keyword based search is having prime importance. In this paper some dispute like Diverse Data Models, Query Forms: Complexity versus Expressive Power , Search Quality Improvement , Evaluation are also discussed [1]

Keyword search is one of the powerful method for information retrieval. One of the advantage of keyword searching query is the simplicity users do not have to learn a complex query language, and can issue queries without any prior knowledge about the structure of the underlying data. Since the keyword search query interface is very flexible, queries may not always be precise and can potentially return a large number of query results. XML search return nested XML element contain desired keywords. Second, the nested structure of

XML implies that the approach of ranking is no longer at the granularity of a document, but at the granularity of an XML element. In this paper, we present the XRANK system for handling features of XML search. XRANK generalizes HTML search engine such as Google[**2**]

For improve the efficacy of keyword search in XML data smallest lowest common ancestor (SLCA) semantics is used . A keyword search using the SLCA semantics returns nodes in the XML and it satisfy conditions: (1) the subtrees rooted at the nodes consist of all the keywords, and (2) the nodes do not have any convenient descendant node that suit condition. To generalize search paradigm to support keyword search ahead the traditional AND semantics to include both AND and OR Boolean operators as well. We first test the properties of the LCA computation and pro- pose improved algorithms to solve the traditional keyword search problem [3]

Here propose two efficient algorithms, Indexed Lookup Eager and Scan Eager, for keyword search in XML documents according to the SLCA semantics. Both algorithms produce part of the answers quickly so that users do not have to wait long to see the first few answers.[4]

Top-k keyword search over probabilistic XML data, which retrieve highest probable k SLCA results. And then two efficient algorithms are proposed. The first algorithm PrStack can catch k SLCA results with the k highest probabilities by scanning the admissible keyword nodes only once. To further advance the efficiency, we propose a second algorithm Eager TopK based on a set of pruning properties which can immediately prune unsatisfied SLCA candidates.[5].

Maximal Marginal Relevance(MMR) criterion diminish relevancy while preserving query relevance in re-ranking retrieved documents and in selecting convenient passages for text summarization. opening results indicate some benefits for MMR diversity ranking in document retrieval and in single document summarization. [6] Diversify search results using temporal expressions (e.g., 1980s) from their contents. Based on pseudo relevant document first identify time intervals of interest to the given keyword query . It then re-ranks query results so as to maximize the coverage of identified time intervals. Using history-oriented queries and encyclopedic resources we show that our method is able to present search results diversified along time.[7]

This paper also uses greedy approach. Different datasets are considered in this to get approach tested thoroughly and relevant document in terms of search result is expected as search result [8]. For structured Search result Differensation Efficient algorithms are designed. that is single swap and multi swap algorithms. To present the applicability of our approach, we implemented a system XRed for XML result discrimination. Degree of difference is count so that it represents the accuracy of search result [9]

III. Proposed System

The issue of diversifying keyword search is primarily studied in IR community. Most of them execute diversification as a post-processing or reranking step of document retrieval based on the analysis of result set and/or the query logs. When we give a query q first system identify candidate keywords of that query with the help of feature selection. After that system validates and checks the quality of fetched candidate keywords for searching purpose with the help of XML keyword search diversification model. Here a method is used called anaphora method that is use of an expression whose explanation depends upon another expression in context . In a narrower sense, anaphora is the use of an expression that depends specifically upon antecedent. Two efficient algorithms are used baseline algorithm and anchor based pruning algorithm. Baseline algorithm retrieve diversified keyword search result .Anchor based pruning improve efficiency of keyword search diversification.

First user query is considered and searching keywords are found. Then concluding the searching keywords of user system used by mutual information model and then calculate the correlation values so that we will get new keyword query. Concluding the mutual information amongst the keywords , their context based relevant keywords or featured term for new query is searched over XML dataset Original keywords and fetched keywords has some common information hence their relevance factor is calculated. After relevance factor calculation their novelty factor is calculated. It provide top –k results

IV. Conclusion

This paper presented an approach to search diversified results of keyword query from XML data based on the contexts of the query keywords in the data. for effectiveness propose another method called anaphora method the method. This method help to reduce time taken for search process and also it reduce computational overheads.

This new approach that we proposed help to effectively Compute new SCLA results. Diversification algorithm can returned qualified search intention and result to user in short time. we designed three efficient algorithms based on the observed properties of XML keyword search results. varied the effectiveness of our diversification model by analyzing the returned search intentions for the given keyword queries over DBLP data set.

References

[1].   Y. Chen, W. Wang, Z. Liu, and X. Lin, "Keyword search on struc- tured 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. Carbon ell and J. Goldstein, "The use of MMR, diversity- based re-ranking for R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," in Proc. 2nd ACM Int. Conf. Web Search D ata Mining, 2009, pp. 5–14.

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

[8].   Z. Liu, P. Sun, and Y. Chen, "Structured search result differ- entiation," J. Proc. VLDB Endowment, vol. 2, no. 1, pp. 313–324, 2009.

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