Survey on Approaches and Methodologies of Keyword Search Diversification Model Over XML Data

Nita Bankar [1], *A.N.Nawathe* [2]

[1] *Student, Computer Department, Amruthvahini College Of Engineering, Maharashtra, India.*
[2] *Assistance Professor, Computer Department, Amruthvahini College Of Engineering, Maharashtra, India.*

ABSTRACT

Keyword search generally used to search large amount of data. There are certain difficulties occurred to answer some queries due to their ambiguity. Also due to short and uncertain keywords diversification of keyword creates a problem. We proposed a system to address these problems. Our system automatically expands keyword search. It is based on different context information of the XML data. Our system firstly selects a feature selection model for designing an effective XML keyword search from a large database. Then it will automatically diversify the keyword search. For searching keyword from XML data a short and vague keyword query is used. Feature selection model is used to derive search candidate of the query search. Our proposed model proves the effectiveness of our system by evaluating real as well as synthetic datasets. In this system more efficiency can be achieved as we proposed pruning algorithm and Hadoop platform for implementation of our system.

Keywords- *XML keyword search, context-based diversification*

1.Introduction

We proposed an approach for searching diversified results of keyword query from XML data based on the contexts of the query keywords in the data. Searching a KEYWORD over structured and semi-structured information enables users to fetch information without the any experienced query languages and database structure Keyword search is most popular because no need to learn query language and database structure. The problem of diversifying keyword search is firstly studied in IR. Most of them perform diversification as post-processing step of document retrieval based on the analysis of result set . In IR, keyword search diversification is designed at the topic or document level.[1] Uncertainty is widespread in many web applications, such as information extraction, information integration, web data mining, etc. The flexibility of XML data model allows a natural representation of uncertain data.[2]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 .Meta is data about data and it is also a method of post-process search result analysis. When the given keyword query only contains a small number of unclear keywords, it would become a challenging problem to derive the search intention due to the high ambiguity or double meaning of this type of keyword queries.The strengths of XML is that it can be used to represent structured facts (records) as well as unstructured facts[3] A recent direction to improve the effectiveness of keyword search in XML data is based on the notion of smallest lowest common ancestor(SLCA). The flexibility of XML data model allows a natural representation of uncertain data.[4] The classical information retrieval scenario consists of identifying a fairly small number of document, that are expected to satisfy a user's information need, in response to a query expressing the need[5]. Diversification aims at minimizing the risk of user's dissatisfaction by balancing relevance and novelty of search results.[6] Result Diversification has been recently introduced for relational datasets as well as recently been examined and applied to several different domains [7].

2 Literature Survey:

Search the result of diversification can be achieved by the relevance of query and documents and similarity between documents in the set of result. The main objective of diversification was to find the optimal set of candidates, which is both relevant and diverse.The relevance and diversification of a search result can be combined by using following different approaches. Often, in the real world, entities have two or more representations in databases. Duplicate records do not share a common key and/or they contain errors that make duplicate matching a difficult task. Errors are introduced as the result of transcription errors, incomplete information, lack of standard formats, or any combination of these factors.[8] Search diversification is different from our context in a couple of fundamental ways. First, the primary motivation for search result diversification is keyword disambiguation[9] We propose a new ranking formula by adapting existing IR techniques based on a natural notion of virtual document. Compared with previous approaches, our new ranking method is simple yet effective, and agrees with human perceptions.[10]

1. Max-sum diversification

.

This first approach use to compute the sums of relevance score of each document with the query that user want to search, it also computes the diversity of each document in the relevant result set. At the end, combine the relevance score and diversity of the query.

2. Max-min diversification

The target of second approach is to increase the sum of those documents which have minimum relevance and maximum dissimilarity within the result set. Max-min diversification is important for those documents which have low relevance and diversity but may be important for the user.

3. Average dissimilarity diversification

This diversification technique is use to sum the original relevance for a result with the average dissimilarity of all documents in the result set. The main theme of average dissimilarity maximizes the sum over the whole set.

4. Max-sum of max-score diversification

This function gives more importance to the relevance between query and documents. The max-sum of max-score produces a set of results that have the maximal relevance sum and then adds maximum diversity into final result set.[11]

3 METHODOLOGY

We can broadly categorize these methods into two groups

1) pre-retrieval

   This method predict the difficulty of a query without computing its results. These methods use the statistical properties of the terms or content in the query to measure *specificity, unclearity,* of the query to predict its difficulty.

2) post-retrieval

This method utilize the results of a query to predict its difficulty[12] Some methods use machine learning techniques to learn the properties of difficult queries and predict their hardness[13].

3. CONCLUSIONS

We proposed a system that searches keyword query search over a big data in XML format. In this project we are contributing a Hadoop platform that will help for analysis of big data in XML format. Therefore our system works efficiently for big data. Also our system can achieve more efficiency as we proposed pruning algorithm and Hadoop platform for implementation of our system. Our proposed pruning algorithm is used for refinement of diversified result.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.

4. REFERENCES

[1] Jianxin Li, Chengfei Liu, Member, IEEE, and Jeffrey Xu Yu, Senior Member, IEEE, Context-Based Diversification for Keyword Queries Over XML Data. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 3, MARCH 2015

[2] Jianxin Li[1], Chengfei Liu[1], Rui Zhou, Wei Wang, Top-k Keyword Search over Probabilistic XML Data,

[3] Pooja Chudiwal1, A. C. Lomte, Diversifies XML Keyword Search Based on its Different Contexts in the XML Data. International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

[4] Jianxin Li[1], Chengfei Liu[1], Rui Zhou, Top-k Keyword Search over XML Data.

[5] Albert Angel, Nick Koudas, Efficient Diversity-Aware Search

[6] Elena Demidova1, Peter Fankhauser1, 2, Xuan Zhou3 and Wolfgang Nejdl11L3S Research Center, Hannover, Germany,Fraunhofer IPSI, Darmstadt Germany,CSIRO ICT Centre, Australia, *DivQ*: Diversification for Keyword Search over Structured Databases.

[7] Mahbub Hasan UC Riverside hasanm@cs.ucr.edu,Abdullah Mueen,UC Riverside mueen@cs.ucr.edu,Vassilis Tsotras UC Riverside tsotras@cs.ucr.edu Eamonn Keogh UC Riverside eamonn@cs.ucr.edu Diversifying Query Results on Semi-Structured Data.

[8] Marcos R. Vieira1, Humberto L. Razente2, Maria C. N. Barioni2, Marios Hadjieleftheriou3, Divesh Srivastava3, Caetano Traina Jr.4, Vassilis J. Tsotras On Query Result Diversification.

[9] Dr.M.Mayilvaganan, M.Saipriyanka Efficient and Effective Duplicate Detection Evaluating Multiple Data using Genetic Algorithm. International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 9, September 2015

[10] Debmalya Panigrahi CSAIL, MIT Cambridge, Atish Das Sarma Gagan Aggarwal Andrew Tomkins Google Research Mountain View, CA, Online Selection of Diverse Results

[11] Adnan Abid, Naveed Hussain, Kamran Abid, Farooq Ahmad,Muhammad Shoaib Farooq,Uzma Farooq, Sher Afzal Khan, Yaser Daanial Khan[1],Muhammad Azhar Naeem, Nabeel Sabir, A survey on search results diversification techniques. Neural Comput & Applic
DOI 10.1007/s00521-015-1945-5

[12] Miss. Varsha Vetal, Mrs. Sanchika Bajpai. Review: Predicting The Efficiency of Difficult Keyword Queries Over Databases, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 11, November 2014.

[13] Shiwen Cheng, Arash Termehchy, and Vagelis Hristidis, Efficient Prediction of Difficult Keyword Queries over Databases, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 6, JUNE 2014.