ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Analysis of Diversification for Keyword Search Queries and Context Based XML Data

Manchala Chalapathi Rao
Pursuing M.Tech,
Dept of CSE,
Chebrolu Engineering College.

B.V.V.S.Prasad
Associate Professor,
Dept of CSE,
Chebrolu Engineering College.

Abstract:

Keyword query is an ordinary users to search vast amount of data, the ambiguity of keyword query makes it difficult to effectively answer keyword queries, especially for huge and vast keyword queries. To meet this challenging problem, our analysis proposes an approach that automatically diversifies XML keyword search based on its different contexts in the XML data. Consider huge amount of keyword query and XML data to be searched, we first derive keyword search candidates of the query by a simple feature selection model. Second identify design an effective XML keyword search diversification model to measure the quality of each candidate. Next, baseline efficient algorithms are proposed to incrementally compute top-k qualified query candidates as the diversified search intentions. Compare 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 on real and synthetic data sets demonstrates the effectiveness diversification model and the efficiency of algorithms.

Keywords:

Data Mining, Search Engine Optimization, XML Dataset, Baseline Algorithm

I. INTRODUCTION:

Keyword search is the most important information Discovery technique because the user does not need to know either a query language or the underlying structure of the data. Large number of techniques is used in XML search system. Keyword search is the technique use for the retrieving data or information. Keyword search can be implementing on machine learning databases, also it possible on graph structure which combines relational, HTML and XML data. Keyword search use number of techniques and algorithm for storing and retrieving data, less accuracy, does not giving a correct answer, require large time for, searching and large amount of storage space for data storage. Data mining or information retrieval is the process to retrieve data from large database and transform it to user in un-derstandable form easily gets that information. One important advantages of keyword search is user does not require a proper knowledge of database queries. User easily inserts a keyword for searching and gets a result related to that keyword. Keyword search on relational databases find the answer of the tuples which are connected to database keys like primary key and foreign keys. So this system also present which comparative techniques used for keyword search like DISCOVER, BANKS, BLINKS, EASE, and SPARK.

Existing techniques for information retrieval on real world databases and also experimental result indicate that existing search techniques are not capable of real world information retrieval and data mining task. Data mining is finding insights which are statistically reliable from data, identification of records which does not match the usual patterns might be interesting that require further investigation. Association searches for relationships various attributes like milk and bread along with jam. So providing a good discount on combination can enhance the sales. Process of grouping together values in the data that have similar patterns but these patterns are not known in advance. Analysing the data we make clusters of employee who reach the target more than ten times per week and other who make less than 10 transactions.

It is the process of grouping the data into different classed on the basis of previously known structures. For example we make classification for example student percentage above 70% as distinction, between 60 to 70% percentage first class and below 60% average. Regression attempts to find a function which models the data with the least error fits the data onto the function so that one value can be derived from another.

SECTION II

2.1. Query based search engines:

Programming search database and internet sites for the documents containing keywords specified by a user, primary function is providing a search for gathering and reporting information available on the internet or a portion of the internet. Communication to the search engines requirements so that can recommend most relevant websites related to search. Searched by the requirement that is being given the text on the page and titles and description that are given. When we use the search engine in relation to the XML usually referring to the actual search forms that search through databases of XML HTML documents available all over. Crawler based search engines are those that use automated software read the information on the actual website.

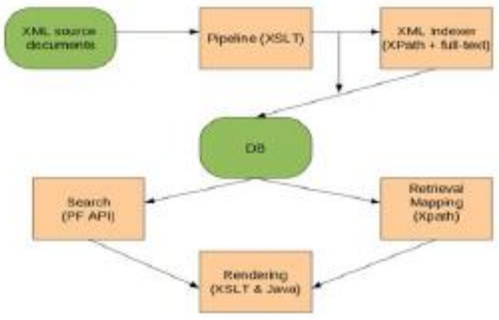


Figure 1 Query based search Engine

This system design has lot of features enabled most of what we needed with a technology stack and we gained a degree of power and flexibility by doing. Beginning the Adhoc query capabilities both MarkLogic and eXist which we had been using provided. When executing queries in a setting with a large amount of data indexes are critical may execute in less than millisecond while the same query could easily take so long that it would effectively never complete. Sometimes users control over the indexes that are generated and how they are used to resolve queries. Provide basic indexes that can be applied automatically and relied on to provide value for a wide range of queries, user gives the clear information about the output of the optimizer tricked by otherwise insignificant syntactic constructs like variables if the user is made aware of this then often rectify the situation by rewriting the queries, allow the user to specify indexes explicitly users can be relied on to know when there are especially interesting sequences to be indexed, provides users with query constructs that references the indexes directly.

2.2. Data Mining Search Engine:

Search Engine Optimization is the procedure of improving the visibility of a website or webpage in search engine unpaid searched results by increasing Search Engine Results Page ranking. Optimization may target different types of search like image search, local search, video search, academic search, new search, industry specific vertical search .It can also be define as the process of affecting the visibility of a website or webpage in search engine. XML is an immense, huge and dynamic data collection that includes infinite hyperlinks and volumes of data usage information-hence requires effective data mining. But huge data is still a challenge in knowledge discovery. Web pages have dynamic data and do not and lack knowledge of internet usage. Hence user gets lost within huge amount of data. A given user generally focuses on only a tiny portion of the Web, dismissing the rest as uninteresting data that serves only to swamp the desired search results.

2.2.1. Keyword-based search:

This includes search which use keyword indices or manually built directories to find documents with specified keywords or topics. e.g engines such as Google or Yahoo.

2.2.2. Querying deep Web sources:

Where information such as amazon.com's book data and realtor.com'srealestate data, hides behind searchable database query forms that, unlike the surface web, cannot be accessed through static URL links.

2.2.3. Random Surfing:

That follows web linkage Pointers

SECTION III

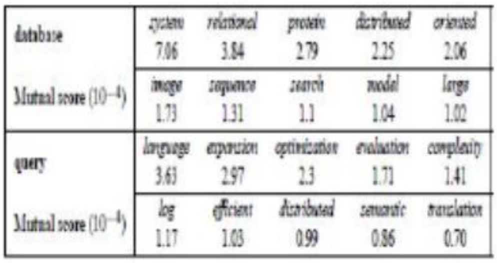
3. Problem Definition:

Now days XML is language which uses to send the messages and to design websites documents. XML keyword performed by searching the exact entered by the user may not result in the effective search user may not know exact keywords. XML data classification and outliers documents may not contain keywords. A search is binary trees may not result in returning documents of AVL trees and Red Black Trees. XML can be effective in this methodology as it can help to classify synonyms related to a keyword.

A keyword query q and an XML data denoted by T, consider a set of possible search intentions Q that are generated by each query to a context using its relevant feature terms in T. the user the top k qualified queries in terms of high relevance and diversification. follow any uniform structure. Web pages contains huge amount of raw data that is not indexed therefore searching in web data has become more complex; time consuming and difficult.

Web not only contains static data but also data that requires timely updating such as news, stock markets, live channels etc. People from different communities have different backgrounds and use internet for different usage purposes. Many have different interests

Table 1 Represents the Search Engine Analysis



database	systems	relational	protein	distributed	oriented
	7.06	3.84	2.79	2.25	2.06
Mutual score (10^{-4})	image	sequence	search	model	large
	1.73	1.31	1.1	1.04	1.02
query	language	expansion	optimization	evaluation	complexity
	3.63	2.97	2.3	1.71	1.41
Mutual score (10^{-4})	log	efficient	distributed	semantic	translation
	1.17	1.03	0.99	0.86	0.70

The different pairs are selected based on mutual information used to criterion for features transformation in machine learning, the variables redundancy feature selection have an XML tree T and its sample result set R(T). Query q is combination of database query over XML dataset shows the mutual information score for the query keywords in q, each one represents in terms of matrix a search intention with the specific semantics query expansion database systems targets to search the publication discussing the problem query expansion in the area of database systems. Query expansion for information. retrieval in Encyclopedia of Database system with relational then the generated query will be changed to search specific query expansion over relational database in which the returned results are empty because no work is reported to the problem to the over relational database.

3.1. Architecture:

Search Engine optimization is procedure of improving the visibility of webpage in search engine natural results by increasing search engine page ranking may target different types of search like image hyperlinks, HTML, XML, video industry search defines asthe process of affecting the visibility of a webpage in search engine. Database is huge and dynamic collection includes highlighting points volumes of data usage information hence requires effective mining is challenge in knowledge discovery. XML pages are more complex than text data do not follow any uniform structure that contains raw data that is not indexed therefore searching in web data has become more complex time consuming and difficult.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

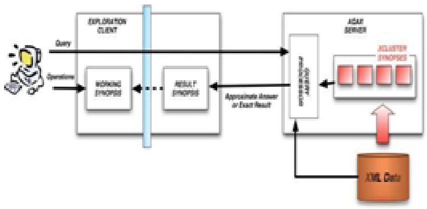
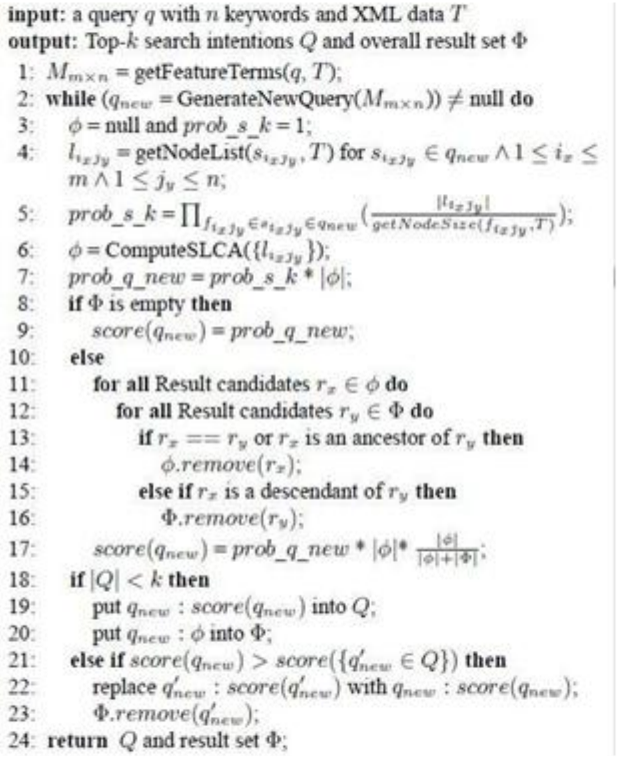


Figure 2 Architecture of XML Search Engine

The procedure of generating a query from the original keyword data to be searched, keyword query q first retrieve the corresponding feature terms for each query keyword and the construct matrix are sorted based on mutual information scores represents a search intention. The aggregated mutual information score of the each search intention represents to some extent the confidence of the context of the query keywords without other knowledge to generate the search intentions and then click the corresponding queries in descending order by aggregated mutual information scores.

20 feature terms for each query keyword and then generate all the possible search intentions from which we further identify the top kqualified and diversified queries the original query. Baseline algorithm retrieves 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 features terms at last compute the SLCAs as keyword search results for each query and measure its diversification score.

Different traditional XML keyword search to detect and remove the duplicated results by comparing the generated results may cover multiple search algorithms to meet the requirement of keyword search diversification.



```
input: a query $q$ with $n$ keywords and XML data $T$
output: Top-$k$ search intentions $Q$ and overall result set $\Phi$
$M_{m \times n}$ = getFeatureTerms($q, T$);
while ($q_{new}$ = GenerateNewQuery($M_{m \times n}$)) $\neq$ null do
    $\phi$ = null and $prob\_s\_k = 1$;
    $l_{i_x j_y}$ = getNodeList($s_{i_x j_y}, T$) for $s_{i_x j_y} \in q_{new} \wedge 1 \leq i_x \leq m \wedge 1 \leq j_y \leq n$;
    $prob\_s\_k = \prod_{f_{i_x j_y} \in s_{i_x j_y} \in q_{new}} (\frac{|l_{i_x j_y}|}{getNodeSize(f_{i_x j_y}, T)})$;
    $\phi$ = ComputeSLCA($\{l_{i_x j_y}\}$);
    $prob\_q\_new = prob\_s\_k * |\phi|$;
    if $\Phi$ is empty then
        $score(q_{new}) = prob\_q\_new$;
    else
        for all Result candidates $r_x \in \phi$ do
            for all Result candidates $r_y \in \Phi$ do
                if $r_x == r_y$ or $r_x$ is an ancestor of $r_y$ then
                    $\phi.remove(r_x)$;
                else if $r_x$ is a descendant of $r_y$ then
                    $\Phi.remove(r_y)$;
        $score(q_{new}) = prob\_q\_new * |\phi| * \frac{|\phi|}{|\phi| + |\Phi|}$;
    if $|Q| < k$ then
        put $q_{new} : score(q_{new})$ into $Q$;
        put $q_{new} : \phi$ into $\Phi$;
    else if $score(q_{new}) > score(\{q'_{new} \in Q\})$ then
        replace $q'_{new} : score(q'_{new})$ with $q_{new} : score(q_{new})$;
        $\Phi.remove(q'_{new})$;
return $Q$ and result set $\Phi$;
```

Worst case all the possible queries in the matrix have the possibility of being chosen as the top-k qualified query candidates, the complexity of the algorithm is generated as

$$O(m^{|q|} * L_1 \sum_{i=2}^{|q|} log L_i)$$

where L1 shortest node list of anygenerated query q is the number of original querykeywords and m is the size of selected features for eachquery keyword, the complexity of the alogirthm can bereduced by reducing the number m of feature terms whichcan be used to bound the number of generated queries.

SECTION IV

4.Comparative Study:

The problem of diversifying keyword search is firstly studied in IR community. Most of them perform diversification as a post-processing or re-ranking 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.

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

V.CONCLUSION:

This work is presented a method to search diversified analysis 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, framework is efficient algorithms based on the observed properties of XML keyword search analysis. Our comparative study, demonstrated the efficiency of proposed algorithms by running substantial number of queries over XMark datasets. At the same time, we also verified the effectiveness of our diversification model by analysing the returned search intentions for the given keyword queries over DBLP dataset and search intentions and results to users in a short time.

REFERENCE:

[1] Peshave, Monica, and KamyarDezhgosha. "How search engines work and a web crawler application." Department of Computer Science, University of Illinois, Springfield USA (2005).

[2]Edgar Damian Ochoa: An Analysis of the Application of selected SEO techniques and their effectiveness on Google's search ranking algorithm, California state university, Northridge, May'2012 - csundspace.

[3]Malaga, Ross A. "Worst practices in search engine optimization." Communications of the ACM 51.12 (2008) pp 147-150.

[4]Vinit Kumar Gunjan, Pooja, Monika and Amit: Search engine optimization with Google, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012, ISSN (Online): 1694-0814.

[5]P. T. Chung, S. H. Chung and C. K. Hui, "A web server design using search engine optimization techniques for web intelligence for small organizations", LISAT IEEE Long Island University Brooklyn NY USA, (2012), pp.1-6.