Volume 3, Issue 6, June-2016, pp. 314-320 ISSN (O): 2349-7084

International Journal of Computer Engineering In Research Trends

Available online at: www.ijcert.org

# Context Based XML Data and Diversification for Keyword Search Queries

[1]Mr. RAHUL HON , [2] Mrs. N.SUJATHA

[1] Pursuing M.Tech(CSE)from Jagruti Institute of Engineering and Technology

[2] Associate Professor, Department of Computer Science and Engineering, Jagruti Institute of Engineering and Technology, Telangana State, India.

**Abstract—** In searching process user enter particular candidate searching keyword and with the help of searching algorithm respective searching query is executed on targeted dataset and result is return as an output of that algorithm. In this case it is expected that meaningful keyword has to be entered by user to get appropriate result set. In case of confusing bunch of keywords or ambiguity in it or short and indistinctness in it causes an irrelevant searching result. Also searching algorithms works on exact result fetching which can be irrelevant in case problem in input query and keyword. This problem statement is focused in this system. By considering the keyword and its relevant context in XML data , searching should be done using automatically diversification process of XML keyword search. In this way system may satisfy user, as user gets the analytical result set based on context of searching keywords. For more efficiency and to deal with big data, HADOOP platform is used. 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, Candidate Keyword ,XML Keyword search, feature selection, diversification process.

## 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.

## II. LITERATURE SURVEY

In this by considering the keyword and its relevant context in XML data , searching should be done using automatically diversification process of XML keyword search is the major area of concern [1].

In this for structured and semi-structured data, various state-of-the-art techniques are discussed for keyword search. In this query optimization , ranking phases , top k important query processing is discussed. Different data models such as XML , graph-structured data is discussed. Application of these concepts is also discussed in which keyword based search is having prime importance. In this paper some problems like Diverse Data Models, Query Forms: Complexity versus Expressive Power , Search Quality Improvement , Evaluation are also discussed [2].

XRANK system is discussed in this paper. Ranked search technique over XML data is considered here. In this paper space saving and performance gaining techniques such as index structure and query evaluation are also focused. XRANK can help in searching for HTML as well as XML documents. Disadvantage: For instance, authors have currently taken a document-centric view, where they assume that query results are strictly hierarchical. Index maintenance is major problem for effective search and which is bottleneck area [3].

In this SLCA-based keyword search approach is discussed. Queries called the Multiway - SLCA approach (MS) is helpful to promote the keyword search beyond and old methods like AND / OR. After LCA analysis improved algorithms are put to solve search problems based on keywords [4].

In this Indexed Lookup Eager and Scan Eager, algorithms are discussed. XML search based on keyword according to SLCA semantics is prime topic of discussion and for this these algorithm are used. Instant search result is the beauty of theses algorithm. XKSearch architecture implementation is discussed in it. The XKSearch system inputs a list of keywords and returns the set of Smallest Lowest Common Ancestor nodes [5].

Query and information relevance is calculated so that unnecessary checks are avoided and effective search is achieved. Hence effective text retrieval

and summarization is achieved. The Maximal Marginal Relevance (MMR) achieves the stopping of redundancy. This approach provides very much relevant data in terms of search result to the end user by effectively minimizing the redundancy [6].

In this paper Risk of dissatisfaction of user is major area of concern. To minimize it systematic approach to diversifying results is discussed in it. For this several techniques such as NDCG, MRR, and MAP are discussed in detail in it. A Greedy Algorithm for Diversification used in it. Among the search result user should find most relevant data is the aim of diversification. Also another aim of this paper is to minimize the rank of best fitted result [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].

In this using test collection based on TREC question answering track this paper discussed the framework which achieves novelty and diversity. In this approach document is linked with the relevant information in it. Chunk of information is in this way get attached with document and which is helpful in at time of search. This piece of information is having content as well as document properties. The major drawback of this approach is that unusual features of document may cause judging error. Some raw data related with the document may delay the search result [9].

Using past query and its analysis provides proper direction for diversification. Past query reformulation provides exact query related behavior of user. Client data request, his ranked structure and query is observed and analysed at client side for proper diversified result. Large query logs are analyzed in this paper from search engine [10].

In this single swap and multi swap algorithms are used in this paper. On structured data differentiation of search results is carried out. Degree of difference is quantified so that it represents the accuracy of search result. Features from the search result are traced and this result is prominently considered in calculation [11].

In this by considering query result and its redundancy, new scheme named re-ranking query interpretations is discussed to diversify the search result. For sub-topics and relevance new proposed technique such as propose $\alpha$-n DCG-W and WS-recall is promoted in it. Algorithm named as Diversification algorithm is used in it. For database query search query similar measure and greedy algorithm is used to obtain diversified query interpretation and its relevance [12].

## III.METHODOLOGY

**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 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 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

**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

**Querying deep Web sources:** Where information such as amazon.com's book data and realtor.com's realestate data, hides behind searchable database query formsthat, unlike the surface web, cannot be accessed through static URL links.

**Random Surfing:** That follows web linkage pointers

**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.

## IV.SYSTEM STUDY

### 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.

### DISADVANTAGES 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.

**PROPOSED SYSTEM:**

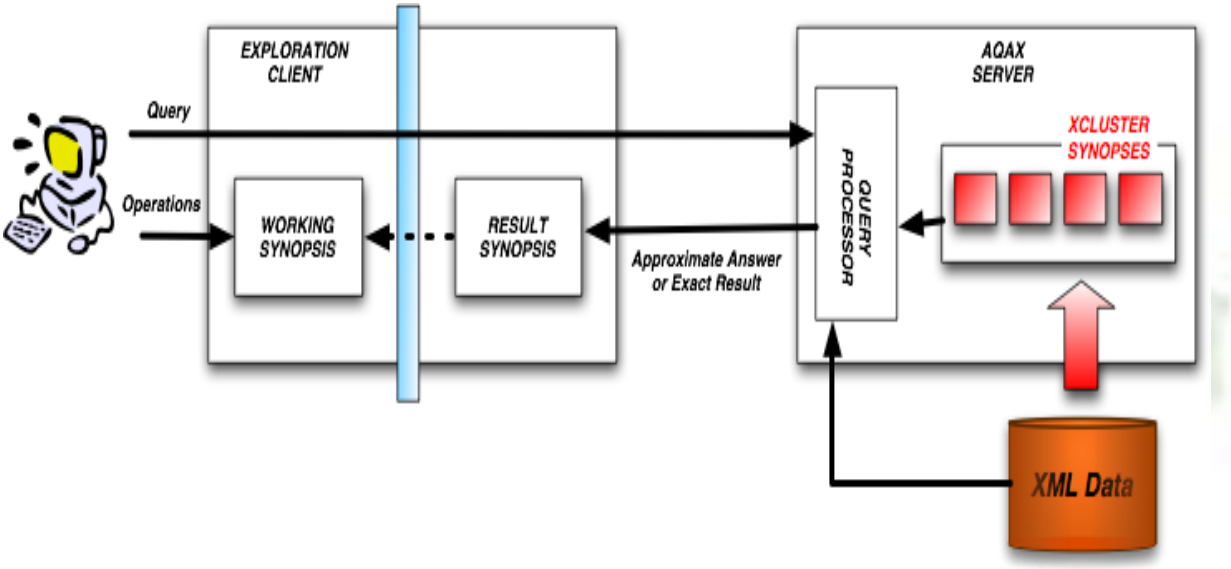
- To address the existing issues, we will develop a method of providing diverse keyword query suggestions to users based on the context of the given keywords in the data to be searched. By doing this, users may choose their preferred queries or modify their original queries based on the returned diverse 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.

**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.

## V. SYSTEM 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 as the 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 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 k qualified 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.

## VI. 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 analyzing the returned search intentions for the given keyword queries over DBLP dataset and search intentions and results to users in a short time.

## REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

[5] F. Radlinski and S. T. Dumais, "Improving personalized web search using result diversification," in Proc. SIGIR, 2006, pp. 691–692.

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

[7] E. Demidova, P. Fankhauser, X. Zhou, and W. Nejdl, "DivQ:Diversification for keyword search over structured databases," inProc. SIGIR, 2010, pp. 331–338.

[8] N. Sarkas, N. Bansal, G. Das, and N. Koudas, "Measure-driven keyword-query expansion," J. Proc. VLDB Endowment, vol. 2,no. 1, pp. 121–132, 2009.

[9] N. Bansal, F. Chiang, N. Koudas, and F. W. Tompa, "Seeking stable clusters in the logosphere," in Proc. 33rd Int. Conf. Very Large Data Bases, 2007, pp. 806–817.

[10] S. Brin, R. Motwani, and C. Silverstein, "Beyond market baskets:Generalizing association rules to correlations," in Proc. SIGMOD Conf., 1997, pp. 265–276.

[11] W. DuMouchel and D. Pregibon, "Empirical bayes screening for multi-item associations," in Proc. 7th ACM SIGKDD Int. Conf.

## ABOUT THE AUTHORS

**Mr.RAHUL HON** is pursuing M.Tech degree in, Computer Science and Engineering from Jagruti Institute of Engineering and Technology, Telangana State, India.

**Mrs.N.SUJATHA** is presently working as Associate Professor in, Department of computer science and engineering, Telangana State,India.She has published several research papers in both International and National conferences and Journals.