# An Effective Method for Searching Keyword Queries Over Semi-Structured Data

**Bodke Nikita[1], Derle Sanjeevani[2], Gite Priyanka[3], Jadhav Priyanka[4], Prof. Kirti Patil[5]**
[1, 2, 3, 4] UG Student, Dept. Of IT., MET's BKC IOE, Nashik, Maharashtra, India
[5] Assistant Professor, Dept. Of IT., MET's BKC IOE, Nashik, Maharashtra, India, India
Email - sanjeevanid014@gmail.com

***Abstract:*** *The search operation of keyword query is challenge for ordinary users to search vast amount of data, the ambiguity of keyword query makes it is difficult to effectively give result to keyword queries. To overcome this complex problem, in this paper we propose an approach that automatically diversifies & search the keywords from XML data. Given small keyword query and XML data to be searched, we firstly access keyword search candidates of the query by feature selection model. And then, we design an effective XML keyword search diversification model for effective keyword searching. After that , baseline solution algorithm are proposed to evaluate the possible generated query candidates representing the diversified search intentions, from which we can find and return top-k qualified query result that are most relevant keyword query.*

***Key Words:*** *keyword search, ranked search, XML data.*

## 1. INTRODUCTION:

To search the information is essential activity of our lives. Web search engines are widely used for searching textual documents, images, and videos. There are also large collections of structured and semi-structured data both comes on web and enterprises, like relational databases, XML, data extracted from text documents, work flows, etc.

In Traditional days,to access the resources, users have to learn structured query languages, they also need to access data of each individual application domain. By creating the databases more search able will increases the information amount that the user may access and also have ability to gain results of searching more efficient as compared to searching of keywords on textual documents, and also increases the usability of databases and make powerful impact on people's lives. Due to huge benefits of supporting keyword search on structured data, it becomes an hot area in database research and development. Various researchers from different departments are joining the workforce to face various challenges in supporting keyword search on structured data.

The main aim of diversification is to minimizing the user's dissatisfaction by balancing relevance of search results. Now a days diversification of search results on unstructured documents is a very big problem, diversification of search results over structured databases has much less attention. Keyword queries over structured data are offering a target for diversification. Single interpretation of a keyword query can't satisfy the users, and there may possibility that multiple interpretations may overlapping the results. The main challenge here is to give users a quick and efficient result of a keyword query in the available database, to enable user to effectively select the interpretation. For example, a user who issued a keyword query "Orange" so we can say that may be he is interested in fruit or the color. On the other hand in document search, where data instances have to be retrieved and analyzed, rich database structures gives more direct and proper way of diversification. For instance, if keyword "Orange" occurs in two database attributes, such as "Fruit" and "Color", each of these presence of keyword can be viewed as a interpretation of keyword with different attributes which offers complementary results. In a database the query disambiguation is performed before the query execution, the computational time for retrieving and filtering redundant search results can be avoided. In the final step the database system calculates only the top-ranked query results to retrieve most relevant and diverse results. Perform searching of queries over semi-structured data.

In early days, a number of schemes have been proposed [1,2] for diversifying results of document retrieval. Several evaluation schemes for result diversification have also been introduced. Most of the techniques perform diversification as a post-processing or ranking step of document retrieval. These methodology first retrieve mutual results and then re-order the result list to achieve diversification. However, this approach can hardly be applied to structured databases. Similar to result re-ranking, clustering is usually performed as a post processing step, and is computationally expensive. Moreover, it makes results difficult to understand to users. To reduce the existing system problem, the diversification problem in XML keyword search, which can directly compute the diversified results without retrieving all the relevant candidates. Given a keyword query, first derive the correlated feature terms for each query keyword from the XML data based on the mutual information in probability manner, which has been used as a criterion for feature term selection. The selection of feature terms is not limited to the labels of XML elements. Each

combination of the feature terms and the original query keywords represents one of diversified contexts that express specific search intentions. It produces top k relevant results.
Application of proposed system:

Database Selection: This method uses techniques that summarize underlying databases by a keyword relationship graph, and select the most relevant data sources with respect to a user keyword search based on derived queries.
Query Generation: This uses techniques that allow a casual user to author new query templates and Web forms by posing keyword searches. The keyword searches are matched against source relations and their attributes to create multiple ranked queries linking the keyword matches. The set of queries is attached to a Web query form, which can be reused by anyone with related information needs. Analytical Processing:To search Keywords on structured and semi-structured data has attracted much research interest area recently, as it allows users to retrieve information from XML data without necessary to learn special query languages and database structure. As compared with keyword search methods in Information Retrieval (IR) that to find a list of related documents, keyword search approaches in structured and semi-structured data concentrate more on specific information contents. In general, When the user's query contains more keywords ,it is easier to get the search intention of user in order to query can be identified. However, when the given keyword query only contains a small number of complex keywords, it would become a very challenging problem to get the user's search intention due to the high ambiguity of this type of keyword queries. Sometimes user involvement is helpful to identify search intentions of keyword queries, a user involvement may be time consuming when the size of relevant result set is large. To overcome this, a method of providing diverse keyword query suggestions to users based on the context of the given keywords in the data to be searched.

## 2. EXISTING SYSTEM:

Now a days, a number of techniques have been implemented for diversifying results of document retrieval for the large text. In paper [3,4,7] author provides Most of the techniques does diversification re-ranking step of document retrieval. These techniques first retrieve co-related results and then arrange it in top-ranked order in ascending or descending order of the result list to get diversification[4]. However, this approach can mainly be applied to structured databases, where retrieval of all relevant data is usually computationally similar to re-ranking of query results, clustering is usually performed as a post processing step, and is computationally expensive. It makes results more complex to understands to end users. recent approaches to database keyword search translate a keyword query into a ranked list of structured queries[1,6,7].Recent approaches to database keyword search translate a keyword query into a ranked list of structured queries, also known as query interpretations. However, the previous query disambiguation approaches consider only the similarities of different query interpretations rather than their diversity. User will get the result which is not related to his need. Query Processing on Semi-structured Data (i.e XML documents) has been addressed in several occasions and there are different types of algorithms have been proposed[1]. diversifying search results is also a well addressed area of research[4]. In paper [4] author provides a general framework for the result diversification problem. Specialized solutions for relational and web databases are also proposed.

## 3. PROBLEM DEFINITION:

It compute the diversified results without retrieving all the relevant candidates[4,7]. First derive the correlated feature terms for each query keyword from the XML data based on the mutual information for feature term selection[8]. The selection of feature terms is not limited to the labels of XML elements. Each combination of the feature terms and the original query keywords represents one of diversified contexts that express specific search intentions. It produces top k relevant results.

## 4. IMPLEMENTATION:

a. Home

Home module does the following process,
-Enter Query
User enter the query for the efficiently search purpose.
-Request for relevant top-k Result
User request to the system for getting relevant top-k results from system.

-Get Result
User will get top-k qualified results from system

b. Keyword Search

XML Keyword Search does the following process,
-Retrieve request from user
The user sends request to system and system retrieves the user request.
-Left to Right and Top to Bottom Scanning
The dataset starts scanning from left to right and top to bottom for the given entered query. It will scan the XML data and produces result if it present in XML data.
-Calculate Mutual Information
Mutual information of two variables is a measure of the mutual dependence .between two variables.
-Store it in Matrix
Given keyword query q with n keywords, first load it's pre-computed relevant feature term from the term correlated information in XML data T. Which is used to construct the matrix Mm*n.

c. Result

In this, it search Result the queries using different models,
-Get Feature Terms
It will get the feature term from the calculated mutual information.
-Calculate MI Score
Generation of new query candidates are in the descending order of their mutual information score.
-Generate New Query
It generates new query candidates q-new from the matrix Mm*n by calculating diversification score.
-Generate Top K Query
After all this process it will return the top k generated query candidates with high relevant to the entered query.

## 5. SYSTEM ARCHITECTURE:

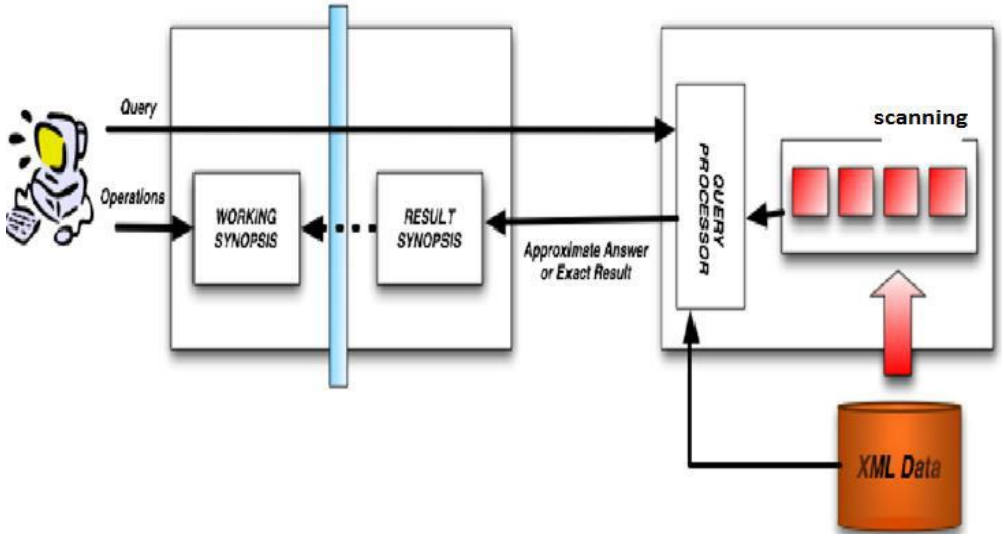


Fig.5.1 Query Scanning Process

## 6. ALGORITHMS:

***6. 1.Baseline Solution:*** In our proposed system different from traditional XML keyword search, our work needs to evaluate multiple expected query candidates and generate a perfect result set, in which the results should be diversified and separated from each result .

Given a keyword query, the intuitive idea of the baseline algorithm is to first retrieve the relevant feature terms with high mutual scores from the term correlated graph of the XML data ;

1.Feature Selection Model
prob (x,T) Be the probability of term x appearing in R(T).
prob (y,T) Be the probability of term x appearing in R(T).
prob (x,y,T) is R(x,y,T) in R(T).

$$MI(x,y,T) = Prob(x, y, T) * \log \frac{Prob(x, y, T)}{Prob(x, T) * Prob(y, T)}$$

### *6.2.Diversification Model*

$$Score(q\ new) = Prob(q\ new\ \ q, T) * DIF(q\ new, Q, T)$$

As such, the top-k diversified query candidates and their corresponding results can be chosen and returned. Different from traditional XML keyword search, our work needs to evaluate multiple expected query candidates and generate a whole result set, in which the results should be diversified and distinct from each other. Therefore, we have to detect and remove the duplicated or unrelated results that have been seen when we obtain new generated results.

## 7. CONCLUSION:

It is 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 were measured by feature selection term relevance to the original query and the novelty of their results. It gives the effectiveness of diversification model by analyzing the returned search intentions for the given keyword queries. From the results,we get proposed diversification algorithms can return qualified search intentions and results to users in a short time.

## ACKNOWLEDGEMENT:

We all are thankful to HOD of our department Prof. Namita Kale for their timely and valuable support.

## REFERENCES:

1. Y. Chen, W. Wang, Z. Liu, and X. Lin, “Keyword search on structured and semi-structured data,” in SIGMOD Conference, 2009, pp. 1005–1010.
2. C. O. Sakar and O. Kursun, “A hybrid method for feature selection based on mutual information and canonical correlation analysis,” in ICPR, 2010, pp. 4360–4363.
3. N. Sarkas, N. Bansal, G. Das, and N. Koudas, “Measure-driven keyword-query expansion,” PVLDB, vol. 2, no. 1, pp. 121–132, 2009.
4. R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, “Diversifying search results,” in WSDM, 2009, pp. 5–14.
5. M. R. Vieira, H. L. Razente, M. C. N. Barioni, M. Hadjieleftheriou, D. Srivastava, C. Traina J., and V. J. Tsotras, “On query result diversification,” in Proc. IEEE 27th Int. Conf. Data Eng., 2011, pp. 1163–1174.
6. M. Hasan, A. Mueen, V. J. Tsotras, and E. J. Keogh, “Diversifying query results on semi-structured data,” in Proc. 21st ACM Int. Conf. Inf. Knowl. Manag., 2012, pp. 2099–2103.
7. E. Demidova, P. Fankhauser, X. Zhou, and W. Nejdl, “DivQ: Diversification for keyword search over structured databases,” in Proc. SIGIR, 2010, pp. 331–338.
8. H. Peng, F. Long, and C. H. Q. Ding, “Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, Aug. 2005.