**Context-Based Diversification for Keyword Queries Over XML Data**

While keyword query empowers ordinary users to search vast amount of data, the ambiguity of keyword query makes it difficult to effectively answer keyword queries, especially for short and vague keyword queries. To address this challenging problem, in this paper we propose an approach that automatically diversifies XML keyword search based on its different contexts in the XML data. Given a short and vague keyword query and XML data to be searched, we first derive keyword search candidates of the query by a simple feature selection model. And then, we design an effective XML keyword search diversification model to measure the quality of each candidate. After that, two efficient algorithms are proposed to incrementally compute top-k qualified query candidates as the diversified search intentions. Two 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. At last, a comprehensive evaluation on real and synthetic data sets demonstrates the effectiveness of our proposed diversification model and the efficiency of our algorithms.

**EXISTING SYSTEM:**

1. 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.
2. 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 these 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:**

1. 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.

2. 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.
3. 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.
4. 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.
5. 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:

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

## SYSTEM SPECIFICATION

### Hardware Requirements:

- System : Pentium IV 3.5 GHz.
- Hard Disk : 40 GB.
- Floppy Drive : 1.44 Mb.
- Monitor : 14’ Colour Monitor.
- Mouse : Optical Mouse.
- Ram : 1 GB.

### Software Requirements:

- Operating system : Windows XP or Windows 7, Windows 8.
- Coding Language : Java – AWT,Swings,Networking
- Data Base : My Sql / MS Access.
- Documentation : MS Office
- IDE : Eclipse Galileo
- Development Kit : JDK 1.6