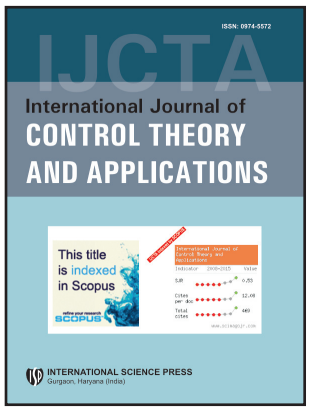

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 26 • 2017

A Novel Mechanism for Information Retrieval with XML Keyword Search

J. Rajanikanth[a], R. Shiva Shankar[b], V.V. Sivarama Raju[c] and Chinta Someswara Rao[d]

[a-d]*Department of CSE, S.R.K.R Engineering College, Bhimavaram, W.G. District, Pin-534 204, A.P. India*

Abstract: Information retrieval plays a vital role in present digital world because of the usage and need of the internet users. For information retrieval some mechanism are found in the literature but there is a scope for newer mechanisms. For this purpose, in this paper we proposed a new information retrieval mechanism with XML keywords on its different contexts of XML data.

Keywords: Information retrieval, XML, Keyword search, internet users.

1. INTRODUCTION

In information retrieval (IR) search strategies, it prefer to listing of pertinent records in organized and semi organized information[1-3]. An approach has been developed, which makes simple to users for searching the content within the given information. This issue of diversifying keyword search is initially examined in IR community huge numbers of them perform diversification like a publish-processing or re-ranking step of document retrieval in line with the analysis of result set. In keyword search diversification, it is fundamental to consider both organized information and semi organized information. So consequently, an appropriate review has been instated from the diversification condition in XML keyword search, which specifically figures the diversified results without retrieving all of the relevant candidates[4]. For this purpose according to mutual information, the co-related feature terms for every query keyword have been inferred as qualifying basics for feature selection.

To identify appropriate results we have to first identify keywords in query. Then for each keyword extract the correlated feature terms keywords from a given XML data set based on predefined metadata and its probabilistic features [5, 6]. This process is similar to the feature selection. The selected feature terms are not same as the labels of XML elements. Each individual combination of the feature terms and query keywords may represents one of diversified contexts. After analyzing the context of diversified query in terms of its relevance with original query and novelty of produced result, will get appropriate queries. To work with large xml data T, our basic aim is to derive top-*k* expanded query candidates from a given query Q with more relevance and maximal diversification where every candidate in candidate list represent the search intention of *q* in T. To efficiently figure diversified

keyword search, two improved algorithms has been exhorted and along with that a Fundamental Multitudinous SLCA framework been proposed which is more proficient.

In currently used HTML based search engines, HTML is a presentation language and is not able to capture semantics[7]. XML allows for extensible element tags which can capture additional semantics. It is a simultaneously human and machine-readable format and supports Unicode, allowing almost any information in any written human language to be communicated. Also XML can represent the most general computer science data structures: records, lists and trees. In HTML data can be only stored but cannot transfer. But in XML data can be stored as well as transferred, while doing this we can make any further enhancement to the code according to the user requirement and get the expected output with user friendly tags. XML is widely used and standard format.

1.1. Motivations

In IR, keyword search diversification is designed at the topic or document level. For e.g., Aggarwal et. al., model user intents at the topical level of the taxonomy and obtain the possible query intents by mining query logs[8]. However, 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. To improve the precision of query diversification in structured databases or semi structured data, it is desirable to consider both structure and content of data in diversification model. So the problem of keyword search diversification is necessary to be reconsidered in structured databases or semi structured data. Liu et. al., is the first work to measure the difference of XML keyword search results by comparing their feature sets[9]. However, the selection of feature set in is limited to metadata in XML and it is also a method of post-process search result analysis. Different from the above post-process methods, another type of works addresses the problem of intent-based keyword query diversification through constructing structured query candidates their brief idea is to first map each keyword to a set of attributes (metadata), and then construct a large number of structured query candidates by merging the attribute-keyword pairs. They assume that each structured query candidate represents a type of search intention, i.e., a query interpretation. However, these works are not easy to be applied in real application due to the following three limitations:

- 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;
- Similar to, the process of constructing a structured query has to rely on the metadata information in XML Digital Bibliography and Library Project dataset.

To address the above limitations and challenges, a formal study of the diversification problem in XML keyword search has been initialized, which can directly compute the diversified results without retrieving all the relevant candidates. Towards this goal, given a keyword query, first derive the co-related feature terms for each query keyword from XML Digital Bibliography and Library Project dataset based on mutual information in the probability theory, which has been used as a criterion for feature 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 may represent one of diversified contexts (also denoted as specific search intentions). And then, 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, one baseline algorithm and two improved algorithms based on the observed properties of diversified keyword search results has been proposed. Given a keyword query q and an XML Digital Bibliography and

Library Project dataset T, our target is to derive top-k expanded query candidates in terms of high relevance and maximal diversification for *q* in T. Here, each query candidate represents a context or a search intention of *q* in T.

2. LITERATURE SURVEY

G. Aggarwal et. al., in [10], proposed the phenomenal growth in the volume of easily accessible information via various web-based services has made it essential for service providers to provide users with personalized representative summaries of such information. Further, online commercial services including social networking and micro-blogging websites, e-commerce portals, leisure and entertainment websites, etc. recommend interesting content to users that are simultaneously diverse on many different axes such as topic, geographic specificity, etc. The key algorithmic question in all these applications is the generation of a succinct, representative, and relevant summary from a large stream of data coming from a variety of sources. In this the formal model is optimization problem, identify its key structural characteristics, and use these observations to design an extremely scalable and efficient algorithm. Analyze the algorithm using theoretical techniques to show that it always produces a nearly optimal solution.

The most relevant work in [11] where Demidova et. al., first identified the attribute-keyword pairs for an original keyword query and then constructed a large number of structured queries by connecting the attribute-keyword pairs using the data schema (the attributes can be mapped to corresponding labels in the schema). The challenging problem is that two generated structured queries with slightly different structures may still be considered as different types of search intentions. However, in diversification model the work utilized to represent different query suggestions and the feature terms are selected based on their mutual correlation and the distinct result sets together. The structure of data is considered by satisfying the exclusive property of SLCA semantics.

In **[**12**],** M. Kolla et. al., use test collection based on TREC question answering 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 getting 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

Different datasets are considered in [13], by Xiangfu Meng et. al.**,** to get approach tested thoroughly and relevant document in terms of search result is expected as search result. User generated content has been fueling an explosion in the amount of available textual data. In the context, it is also common for users to express, either explicitly (through numerical ratings) or implicitly, their views and opinions on products, events, etc. This wealth of textual information necessitates the development of novel searching and data exploration paradigms. However, in contrast to faceted search which utilizes domain-specific and hard-to-extract document attributes, the refinement process is driven by suggesting interesting expansions of the original query with additional search terms. The query-driven and domain-neutral approach employs surprising word co-occurrence patterns and (optionally) numerical user ratings in order to identify meaningful top-*k* query expansions and allow one to focus on a particularly interesting subset of the original result set. The proposed functionality is supported by a framework that is computationally efficient and nimble in terms of storage requirements. The solution is grounded on Convex Optimization principles that allow us to exploit the pruning opportunities offered by the natural top-*k* formulation of our problem. The performance benefits offered by our solution are verified using both synthetic data and large real data sets comprised of blog posts.

The exploration work in [14] by Y. Li, C. Yu, and H.V. Jagadish, proposed SFQ interface because the clients may have just restricted information about XML structure and thus not able to deliver a right XQuery. Thus, by utilize keyword-based search by presenting thought of importance lowest common anchor for finding related nodes in XML report. Developing interfaces to enable casual, non-expert users to query complex structured data has been the subject of much research over the past forty years. Since such interfaces allow users to freely query data without understanding its schema, knowing how to refer to objects, or mastering the appropriate formal query language, called as schema-free query interface. However, schema-free query interface systems are challenged by three hard problems. First, there still lack a practical interface. Natural Language Interface (NLI) is easy for users but hard for machines. NLP techniques of today are still not reliable to parse out the relational structure from natural language questions. Keyword query interface, on the other hand, has limited expressiveness and ambiguity inherited from the natural language terms used as keywords. Second, people have many different ways to express or model the same meaning, which can result in the vocabulary and structure mismatches between the user's query and the machine's representation. This is often referred to as the semantic heterogeneity problem Third, the Web has seen increasing amounts of open domain semantic data with heterogeneous or unknown schemas, for this a new schema-free query interface is introduced and called as SFQ interface, in which the user explicitly specifies the relational structure of the query as a graphical "skeleton" and annotates it with freely chosen words, phrases and entity names. By using SFQ interface, users can work around the unreliable step of extracting complete relations from natural language queries.

S. Cohen et. al., in [15], tells about the proximity which is included in the ranking formula in terms of the size of the relationship tree and thus, it is not affected by the order of children. XSEarch employs more information-retrieval techniques. The main contribution is in laying the foundations for a semantic search engine over XML documents. XSEarch returns semantically related fragments, ranked by estimated relevance. Our system is extensible, and can easily accommodate different types of relationships between nodes. We have shown that it is possible to combine these qualities with an efficient, scalable and modular system. Thus, XSEarch can be seen as a general framework for semantic searching in XML documents.

3. METHODOLOGY

To achieve the highest possible level of efficiency while using the Information Retrieval (IR) system and highest possible level of accuracy while querying results from IR. To ensure that the input is acceptable and understood by the user

When the demonstration starts, the theme of the project is in the final stages while viewing web results. Then applications output data will only provide a likely overview of real-world transactions involving client-server architectures. Methods to increase the accuracy of output data include: repeatedly performing operations like querying, searching and comparing results, dividing events into batches and processing them individually, and checking that the results of these operation are cohesive and stable

To effectively answer the users query by typing a keyword in a search engine. In proposed method main focus is on increasing the efficiency of the searching process by using the Mslca technique, i.e., according to the given query the system should suggest the best auto suggestions and the buffer level minimizes when compare to the existing system.

Existing system: Efficient query retrieval systems are implemented for RDBMS systems only and not for XML based systems. A user tries to compose a keyword query, while being suggested auto completions from servers with respect to their keywords. Even though this concept is nothing new for RDBMS based systems, this is a new information-access paradigm for XML based systems.

For suggestions application of a baseline solution (BE) is vital, although most times the suggestions might not be useful prompting to explore better systems. The BE do not support users expanded knowledge domains. And then, two anchor-based pruning algorithms are designed to improve the efficiency of the keyword search diversification by utilizing the intermediate results. Query results are not supported by in-cohesive keywords that are not present in DBLP dataset.

So a better system is required that supports users expanded knowledge domains and also robust to minor errors in keywords.

Proposed system: In order to improve the quality of auto suggestions on XML data, a baseline algorithm is used to retrieve the diversified keyword search results. And then, two anchor-based pruning algorithms are used and along with that a fundamental multitudinous SLCA algorithm has been proposed. These are designed to improve the efficiency of the keyword search diversification by utilizing the intermediate results and also the efficiency can be obtained by selection of anchor nodes which is in an iterative process. An iterative pruning method over XML Digital Bibliography and Library Project dataset is used to improve relevancy. Here, the system searches in XML Digital Bibliography and Library Project dataset on the fly as the user types in query keywords.

- Benefits of the proposed system includes the following
    - Efficient Auto complete features
    - Supports XML Digital Bibliography and Library Project dataset of varying sizes
    - The minimum buffer level to initiate similarity check once the querying should be began for rendering.
    - The proposed algorithm converts html into xml format which can be used to store and transfer data. This is done by using reverse engineering.
    - Effective index structures and searching algorithms over XML drives top-k results
- Uses the following algorithms and techniques
    - Smallest Lowest Common Ancestor (SLCA) check for relevancy
    - Base Line solution (BE), Pruning Methods (AE, ASPE)
    - Fundamental Multitudinous SLCA
- Produces high search efficiency and result quality over XML Digital Bibliography and Library Project dataset storages.

For keyword query searching only partial results are obtained even though by removing unqualified SLCA results in which anchor based pruning algorithm is used and this results will be in the form of html. In html the data can be stored but not transfer. But by using XML format, the data can be stored as well as transfer also. Therefore we need to convert the html format into xml format. This can be done by introducing Fundamental Multitudinous SLCA framework by using reverse engineering concept. Meanwhile, the effective results can be gained by proposed algorithms when compared to other algorithms and gives top k-qualified results to the ordinary users.

3.1. Flow of Process

Flow of the process is shown in Table 1.

Table 1
Flow of the process

STEP 1	First user query is analyzed and searching keywords are traced
STEP 2	After finalizing the searching keywords of user, system uses mutual information model and calculate the correlation values so that it will be easy to get new query keywords.
STEP 3	After finalizing the mutual information amongst the keywords, their context based relevant keywords or featured term for new query is searched over XML dataset
STEP 4	Original keywords and fetched keywords has some common information hence their relevance factor is calculated.
STEP 5	After relevance factor calculation their novelty factor is calculated. This provides diversified results on the basis of context terms or keywords extracted.
STEP 6	Apply MSLCA for further partial matches and also conversion to xml format.
STEP 7	After getting relevant and novelty result set, top – *k* results are defined.

The actual process is shown in Figure 1.

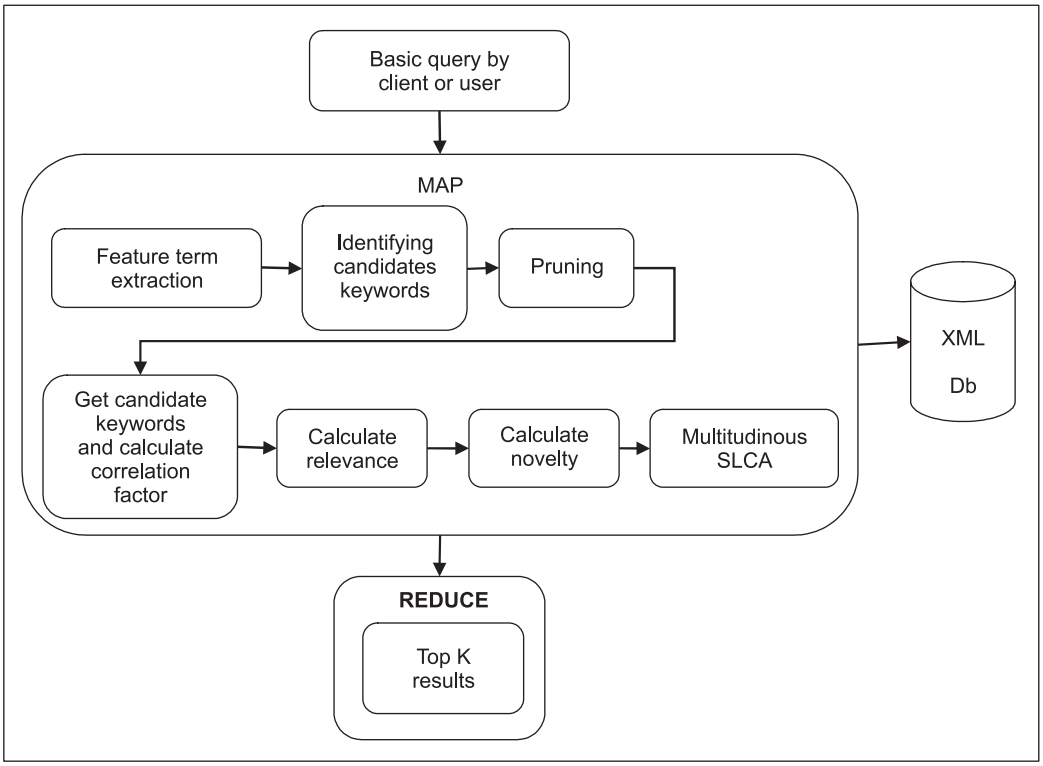

Figure 1: Flow of the process

3.2. List of Modules

The mechanism is divided into the following modules

Baseline Solution: Given a keyword query, the intuitive idea of baseline algorithm is that we first retrieve the pre-computed feature terms of the given keyword query from the XML data T and then we generate all the possible intended queries based on the retrieved feature terms; at last, we compute the SLCAs as keyword search results for each query and measure its diversification score. As such, the top k diversified queries and their corresponding results can be returned to users.

Anchor Based Pruning Solution: By analyzing the baseline solution, we can find that the main cost of this solution is spent on computing SLCA results and removing unqualified SLCA results from the newly and previously generated result sets. To reduce the computational cost, we are motivated to design an anchor-based pruning solution, which can avoid the unnecessary computational cost of unqualified SLCA results (i.e., duplicates and ancestors). In this subsection, we first analyze the interrelationships between the intermediate SLCA candidates.

Anchor-Based Parallel Sharing Solution: Although the anchor-based pruning algorithm can avoid unnecessary computation cost of the baseline algorithm, it can be further improved by exploiting the parallelism of keyword search diversification and reducing the repeated scanning of the same node lists.

Multitudinous-SLCA (M-SLCA): We get top *k* results within short time than the remaining modules which is mentioned. Along with the results, the code will be converted to xml format with the help of reverse engineering concept. By conversion into that format we can not only store the data or information but also we can transfer the data which is not possible in html.

- The mechanism has two output arguments:
  - The representation to be selected for the querying of the next segment
  - The minimum buffer level to initiate similarity check when the querying must be started for rendering:
- Minimizes the page loads by reducing start up delays using the above buffer heuristics mentioned and also supports typo corrections.

In medium.xml generally we have 7900 results. When we perform the search results by giving the keywords to the user it interacts with the server and gives the results. This results were generated by diversification keyword search with the support of BE, AE, ASPE. For this AE and ASPE, slca has been used to lessen the computational cost. Finally result count will be of 290 results which displays on the user screen.

Result analysis: This XML Format is in both XML Properties and XML Nodes form. By changing over to XML form, the information can store as well as transfer which is utilized for further upgrades. The XML results are shown in Figure 2.

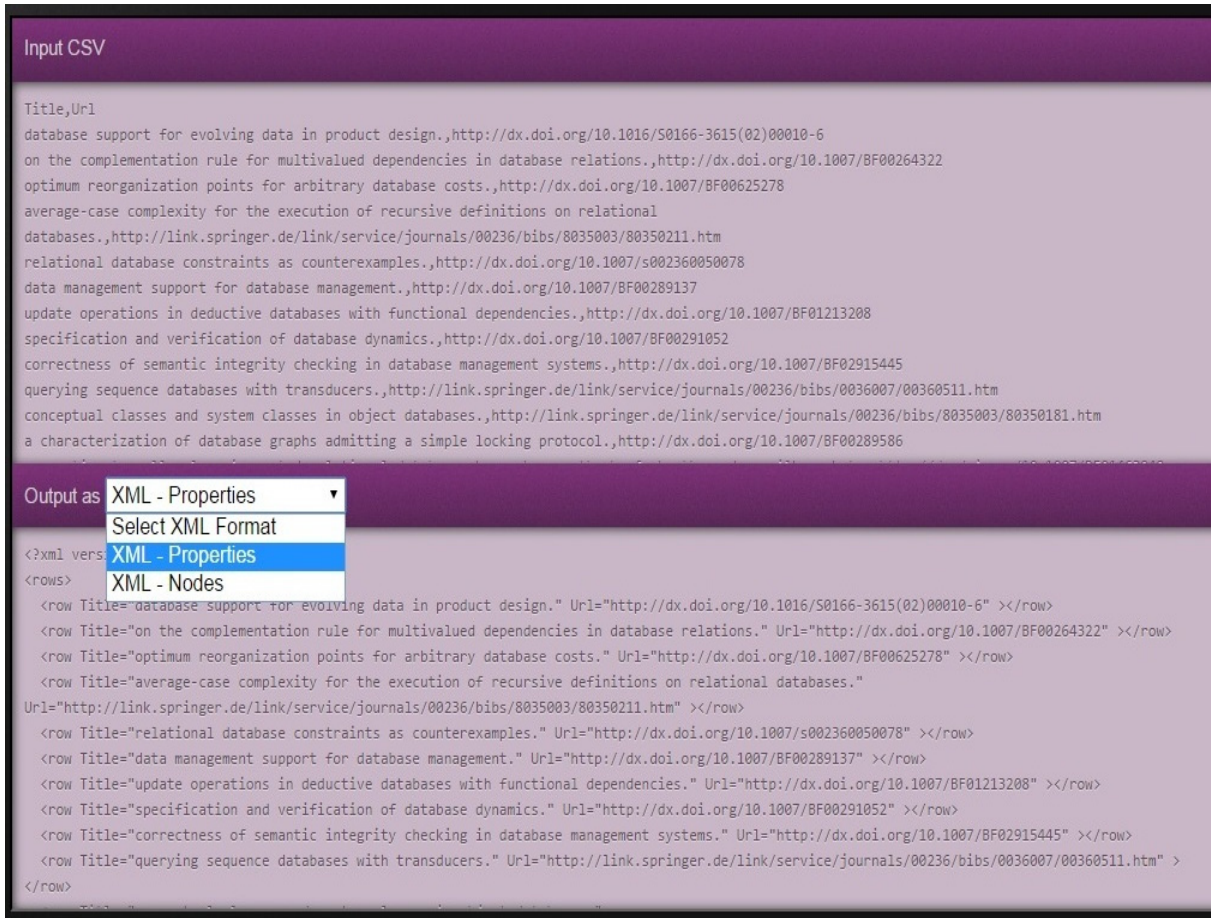

Figure 2: Showing results for converting to XML format

4. CONCLUSIONS

In this paper; we proposed a procedure for search diversified outcomes of keyword query from XML information in accordance with the contexts from the query keywords within the data. From the experimental results it is observed that the proposed mechanism retrieves efficient and more relevant results.

REFERENCES

[1] Datta D, Varma S, Singh SK., "Multimodal retrieval using mutual information based textual query reformulation", Expert Systems with Applications, pp. 81-92, 2017.

[2] Radlinski F, Craswell N., "A theoretical framework for conversational search", In Proceedings of the Conference on Human Information Interaction and Retrieval, ACM, pp. 117-126, 2017.

[3] Chinta Someswara Rao, Dr S Viswanadha Raju, "Concurrent Information Retrieval System (IRS) for large volume of data with multiple pattern multiple (2^N) shaft parallel string matching", Annals of Data Science, Springer, Vol.3, Issue.2, pp..175-203, 2016.

[4] Chinta Someswara Rao, Balakrishna A, Raju MB, Raju SV., "A Frame Work for XML Ontology to STEP-PDM from Express Entities: A String Matching Approach", Annals of Data Science, pp. 469-507, 2016.

[5] Dahak F, Boughanem M, Balla A., "A probabilistic model to exploit user expectations in XML information retrieval", Information Processing & Management, pp. 87-105, 2017.

[6] Ling TW, Zeng Z, Le TN, Lee ML., "ORA-semantics based keyword search in XML and relational databases", IEEE 32nd International Conference on Data Engineering Workshops, pp. 157-160, 2016.

[7] Klusch M, Kapahnke P, Schulte S, Lecue F, Bernstein A., "Semantic web service search: a brief survey", KI-Künstliche Intelligen, pp. 139-147, 2016.

[8] Aggarwal N, Bhatia S, Misra V, "Connecting the Dots: Explaining Relationships Between Unconnected Entities in a Knowledge Graph", Proceedings in International Semantic Web Conference, Springer International Publishing , pp. 35-39.

[9] Liu Y, Nie JY, Chang Y., "Constructing click models for search users", Information Retrieval Journal, pp.1-3, 2017.

[10] D. Panigrahi, A. D. Sarma, G. Aggarwal, and A. Tomkins, "Online selection of diverse results", Proceedings in international conference on Web Search Data Mining, pp. 263–272, 2012.

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

[12] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation", In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 659-666, 2008.

[13] Meng X, Cao L, Zhang X, Shao J., "Top-k coupled keyword recommendation for relational keyword queries", Knowledge and Information Systems, pp.1-34, 2016.

[14] J. Li, C. Liu, R. Zhou, and W. Wang, "Top-k keyword search over probabilistic xml data", Proceedings in IEEE 27th International Conference on Data Engineering, pp. 673–684, 2011.

[15] S. Cohen, J. Mamou, Y. Kanza, and Y. Sagiv., "XSEarch: A Semantic Search Engine for XML", In Proceedings of the 29th international conference on Very large data bases, pp. 45-56, 2003.