ISSN (e): 2250 – 3005 || Vol, 04 || Issue, 8 || August – 2014 ||
International Journal of Computational Engineering Research (IJCER)

XML Retrieval: A Survey

Hasan Naderi[1], Mohammad Nazari Farokhi[2], Nasredin Niazy[3],
Behzad Hosseini Chegeni[4], Somaye Nouri Monfared[5]
[1]*Department of Computer Engineering, Iran University Science and Technology, Tehran, Iran*
[2, 3, 4, 5,] *Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Lorestan, Iran*

ABSTRACT:
Nowadays in the world of the Internet and the Web, great amounts of information in various forms and different subjects are available to users. The available information can be divided into three categories: structured, unstructured and semi-structured. Information retrieval systems traditionally retrieve information from unstructured text which is a text without marking up. XML retrieval is content-based retrieval of structured documents with XML. The aim of XML retrieval is restoring related parts of an XML document that by exploiting the document structure can respond to users' needs. In this research we will examine the XML retrieval. Moreover, models, challenges and retrieve methods exactly are studied.

Keywords: Cas, Co, ComRank, Inex, Information Retrieval, Trex, XML Retrieval

I. INTROUDUCTION:

By rapid development of using extensible language and XML development on the Internet, retrieval of XML data has become one of the most interesting research matters. Since the XML documents are increasingly expanding, engines for search and retrieval can be developed into a set of XML documents in order to perform the search. XML documents have not only textual information, but also contain information about the logical structure of the documents. The logical structure in fact is a tree-like structure that is encrypted by the XML labels. In XML retrieval, elements and components of document are retrieved, not the whole document. Content-based retrieval of XML documents over the past few years has been the most highly regarded which mainly has emerged from the NEXI initiative design [1]. The aim of XML retrieval is restoring related parts of an XML document that by exploiting the document structure can respond to users' needs [2]. Information retrieval systems are often inconsistent with relational databases. In XML retrieval, information needs of users determine as queries, includes key phrases and structured points. Structure, specifies XML elements tracks marked in the set from which system should restore the information [3]. In XML documents and texts, structure and content are separable [4]. An information retrieval system in response to a query returns a ranked list of documents. Then, user examine in the linear case each of them that are in a higher rank [5]. Since the numbers of XML components are generally high, it is necessary that users have systems to retrieve XML, so that components of content have became retrieved and reviewed. One approach could involve the use of summarization that is useful in interactive information retrieval. In interactive XML retrieval, a summary can connect by any one of its document parts which has returned via XML retrieval system [6].

II. THE STRUCTURE OF TEXTUAL INFORMATION

Textual information based on the structure can be divided into three categories:

2.1. Unstructured data: unstructured data means raw text, which through of markings and syntactic labels are separated.

2.2. Structured data: structured data is including data that are already defined. In structured data the user can find out exact and specified respond from their needs.

2.3. Semi-structured data: semi-structured data is between structured data and unstructured data and has stronger structure than unstructured data. We need to incorporate structured information in semi-structured data.

III. INEX

INEX is an international association for the study of XML retrieval. Available approaches of XML retrieval for current structure in ranking and scoring elements are related to returning structures in memory and timing parameters. One approach only returns logical elements such as sections and paragraphs in the search results. Another approach allows users to specify their Structural preferences that consist of structural limitations [7]. INEX can be used in connection with the Xpath to retrieve the XML structural tracks based on what the user specifies in the query. On the other hand, by adding the function about ( ) expands its, which this function is used to filter components [8].

In INEX keywords are combined with structural adverbs. Thus, in response to a question, a ranked list of XML components presents that must contain the following conditions:

1. At least comprising one of the keywords.
2. Also has the considered adverbs [4].

3.1. Our track goals at INEX

In INEX any response is studied to a target. Different track goals are as follows [9]:

- ✓ Adhoc Track
- ✓ Language Processing
- ✓ Interactive Track
- ✓ Multimedia Track
- ✓ Use Case Track
- ✓ Entity Ranking
- ✓ Book Search
- ✓ Link The Wiki
- ✓ Question Answering

IV. CO AND CAS

Users' information needs at INEX are expressed in two ways, CAS and CO. CO approach, shows key phrases based on an approach that is typically used for retrieval information on the Internet. CAS approach, is used a combination of structural and textual marks. In recent years, much work has been done in connection with the CAS that as four sub- tasks was implemented in 2005:

4.1.VVCAS: the target element and limitations of the support elements unclearly were studied.
4.2. SVCAS: limitation of target element explicitly was examined but the limitation of support element is vaguely followed.
4.3. VSCAS: target element and limitation of support element were considered vaguely but limitation of support element is explicitly followed.
4.4. SSCAS: Both limitations of the target element and support element are explicitly considered.
If structural remarks are generated in information needs, in order to demonstrate these two marks, two vague or explicit methods are represented [10]. CAS questions can be solved by analyzing the INEX expressions and decide which indexes used in search. The fundamental ways in analysis CAS questions include the vector space model, DMMS and display XML documents by trees [11]. CO questions are suitable for ordinary users with limited programming skills, and users to achieve the desired information do not need to learn the combination of complex questions from before Xquery and Xpath [12].

4.5. CAS questions are identified in three types:
4.5.1. Routes based questions: route based questions are defined based on Xpath queries such as NEXI.
4.5.2. Clause based questions: clause-centric questions are usually developed from Xquery language.
4.5.3. Parts based questions: sections based questions used XML for the retrieval of XML documents [13, 14].

V. COMRANK SYSTEM IN XML RETRIEVAL :

ComRank system is an Intermediary Search system used for automatic ranking in XML retrieval systems [15]. ComRank system have used a free approach for Intermediary Search so that its results obtained from several systems , its main systems have high ranking ,furthermore its results are achieved from systems which compared with other systems have better operation[16]. Comrank is alike a voting system based on consensus [17].

VI. TREX:

TREX is an XML retrieval system that can use of several summarized structures including the newly defined. TREX can itself manage great but small features, and thus accelerate the assessment of workload to the TOP-K questions .TREX has three methods of comprehensive retrieval, TA and integration. In TREX, summarized structure and reverse lists which are shown in the two tables are stored as the names of elements and sent lists. Evaluation of a NEXI query in the TREX is performed in the two ways of recovery and interpretation [18]. The TREX function in search engine TOPX is generalization of the markup function [19].

VII. RE IN XML RETRIEVAL

Relevant feedback is a technique that allows users to provide feedback on the initial search. The purpose of relevant feedback is that the user's needs express more precisely. RF approaches are proposed to XML retrieval. These approaches by adding extracted words from whole texts and documents enrich questions [20]. In fact, it can be said that relevant feedback is employed to improve results accuracy including extract keywords from documents. In RF, for ranking components from AQR algorithm, separate indexes for each component are created [22].

VIII. EVALUATION OF XML RETRIEVAL

Evaluation of XML retrieval is determined by the member coverage and subject relevance. Member coverage is defined as follow in four ways:

8.1. Exact coverage (E): The principal subject of component is searching for information which components are also.

8.2. Small coverage (S): The principal subject of component is searching for information, but components are not meaningful units of information.

8.3. Large coverage (L): Seeking information on components is presented, but is not the main issue.

8.4. No coverage (N): searching information is not the components subject.

Also, the dimension of subject relevance has four levels which are as follow:

-Highly relevance with the number 3 is specified.

-Relatively relevance with the number 2 is specified.

-Slightly relevance with the number 1 is specified.

-No relevance with the number 0 is specified.

In subject relevance, components are judged in both dimensions and then judgment is combined in a letter - digits code. The composition of relevance coverage is specifying as follows:

$$\mathbf{Q}(rel, cov) = \begin{cases} 1.00 & \text{if} \quad (rel, cov) = 3\text{E} \\ 0.75 & \text{if} \quad (rel, cov) \in \{2\text{E}, 3\text{L}\} \\ 0.50 & \text{if} \quad (rel, cov) \in \{1\text{E}, 2\text{L}, 2\text{S}\} \\ 0.25 & \text{if} \quad (rel, cov) \in \{1\text{S}, 1\text{L}\} \\ 0.00 & \text{if} \quad (rel, cov) = 0\text{N} \end{cases}$$

Formula 1.the compositions of relevant c [23].

2S is a rather relevant part, i.e. it is so small. 2S component provides incomplete information, but answers the question trivially. 3E is a much related component that has much accurate coverage. An unrelated component cannot have precise coverage, so composition of 3N is impossible.

The quantized Q function dose not imposes a dual selection of related / unrelated, and permits to categorize component as low relevance. Some related Components for retrieval set of A are calculated as follows [23].

$$\#(\text{the retrieval relevant cases}) = \sum_{C \in A} \mathrm{Q}\,(\,\mathrm{rel}\,(\mathrm{c}), \mathrm{cor}(\mathrm{c}))$$

Formula 2. relevant components in retrieval set[23].

IX. CHALLENGES IN XML RETRIEVAL:

Challenges in XML retrieval are proposed as follows:

- ✓ parts of the document must be retrieved
- ✓ parts of the document that should be indexed
- ✓ nested element
- ✓ statistical terms
- ✓ heterogeneity model

9.1. Parts of the document must be retrieved

XML retrieval should return the following:

- ✓ parts of documents or XML elements
- ✓ All documents not return

Existing solutions to this challenge is to retrieve documents in a structured way which in fact a system should retrieve the certain part of a document.

9.1. parts of the document that should be indexed

This challenge in the unstructured retrieval, usually is straight, but in structured retrieval has four approaches including of total to detail, of detail to total ,indexing all elements, and lack of interaction in Pseudo-documents . Approach of total to detail is a two-step process that begins with the largest element as a indexing unit, leading to find sub-elements from each element. In this method relevance of a larger element is not necessarily a good predictor of the sub-elements contained in it. The method of detail to total, by considering all of the leaves select the most relevant leaves and expand them into larger units. The approach of indexing all elements, is the most strict approach. In approach of lacking interaction in Pseudo-documents, documents may be meaningless to the user since the units are not contiguous.

9.3. Nested Elements

In this challenge all elements that are small and are not relative leave aside and we keep the elemans which are useful for result.

9.4. Statistical Terms

This challenge has problem in distribution and can be trusted to estimate the frequency distribution of the documents. Calculating the idf term is available solution for the pair of XML documents.

9.5. Heterogeneity Model

It is in two ways of ideal and similar elements in different patterns. In Ideal case, only one model is needed that this model be realized for user. Similar elements are determined in different patterns fall into two different names and different elements in the structure.

X. INDEXING OF XML RETRIEVAL

Several indexing strategies for XML retrieval have been developed as follows:

- ✓ Element-based indexing: allow to each element which based on direct text and generation text, indexing is done. The indexing has one major drawback. Text that appears at the n th logical structure of XML, n-order indexing, thus requiring more index space [24, 25].
- ✓ Only indexing leaf: only allow indexing leaves by the element or elements that are directly related to the leaves.
- ✓ Expanse-axis indexing: text in one continuous element, is used to estimate a statistical expression [26].
- ✓ Selective indexing: includes removal of small elements and selective element type.
- ✓ Distributed indexing: separately for each type of element is created. Ranking model for each indicator separately runs and retrieves a list of ranked elements [22].

XI. RANKING PATTERNS OF XML RETRIEVAL

The ranking patterns are chosen based on indexing strategies and the specific mechanisms, such as expansion and density that at them only leaf elements, are listed. Most of ranking methods create a list of elements with limited or no structural constraints on the associated element in question are ranked.

Distribution or publication of scores for ranking items based on the curve of the leaves is used [27]. Then scores are published upwards to the parent. Ranking model should be applied to each indicator separately and retrieve ranked lists of elements [28].

XII. XML RETRIEVAL MODELS

12.1. Language model

This model combine estimations based on the whole text components and the compact expression components, as well as for improving efficiency and recovery, use from appearance a component in document and main text, and duration of that way.

Sigurbjornsson by using a language model, evaluated different indexing strategies and for retrieving elements created four indicators:

- ✓ Indicator element with traditional overlying elements.
- ✓ Length based on the index, in which the elements of a length pre-set threshold crossed over and are just indexing.
- ✓ Index based on Qrel, where elements specified by heading elements to indexing.
- ✓ Section index, which also indexing other unoverying pages based on structure [30].

12.2. The vector space model (VSM)

Vector space model is the best and most efficient information retrieval models for retrieving unstructured documents [31]. Example of the vector space model is relationship- building tree techniques where the set of document is considered as a tree and documents are under the tree. Question is also a tree, and instead of returning a ranked list of documents from the elements, a ranked list of documents returned [32].

A simple measure of the similarity of the Cq in route search and route of Cd in a document, is the CR similarity function:

$$\mathrm{CR}(c_q, c_d) = \begin{cases} \frac{1+|c_q|}{1+|c_d|} & \text{if } c_q \text{ matches } c_d \\ 0 & \text{if } c_q \text{ does not match } c_d \end{cases}$$

Formula 3.vector space model[32].

Where Cq and Cd are the number of the curves in the search path, and the document path respectively. The final score for a document is computed as a variable of the cosine measure that specifies by SIMNOMERGE, and be defined as follows:

$$\mathrm{SIMNOMERGE}(q,d) = \sum_{c_k \in B} \sum_{c_l \in B} \mathrm{CR}(c_k, c_l) \sum_{t \in V} \mathrm{weight}(q,t,c_k) \frac{\mathrm{weight}(d,t,c_l)}{\sqrt{\sum_{c \in B, t \in V} \mathrm{weight}^2(d,t,c)}}$$

Formula 4. Final score for document[32].

Where V non-structural terms, B the set of all fields of XML and weight (q, t, c) and weight (d, t, c) are the weight of terms t in XML field to searching q and document d.

12.3. Models based on okapi

Let nE element, e = 1,2, ..., nE are the C set. El is the length of element and the avel is the length of average element. Weight for query term j in document d in the collection c, e element is calculated by the following formula:

$$W_j(e,d,c) = \frac{(K_1+1)+f_{e,j}}{k_1\left((1-b)+b\frac{el}{avel}\right)+tf_{e,j}} \log \frac{N-df_j+0.5}{df_j+0.5}$$

Formula5. Formula okapi[9].

Where $tf_{e,j}$ is equal to the frequency of query terms j in element e, df_j is the frequency of documents for query j and N specifies the number of documents [33].

Okapi to calculate the retrieval rate for an element x in a query q using the following formula:

$$Okapi(q,X) = \sum_{j=1}^{q} W_{j,x} \frac{(K_1+1)+f_{x,j}}{k_1\left((1-b)+b\frac{el}{avel}\right)+tf_{e,j}} \times \frac{(k_3+1)qtf_j}{k_3+qtf_j}$$

Formula 6. Okapi for calculating the retrieval rate for an element[9].

Where:

$$W_{j,x} = \log \frac{M-ef_j+0.5}{pf_j+0.5}$$

Where q is the length and ef_j is element frequency of the term j [34].

12.4. Logistic regression model

The relevant probability in any document or document component is estimated according to a series of statistic in a set of document for a series of queries into a series of connected scales to statistics.

The probability P (R | Q, C) to the log-odds of relevance LogO (R | Q, C) can be computed for any two events A and B is a deformation of simple probability P (A | B) / P (A '| B) is as follows:

$$logO(R|Q,C) = b_0 + \sum_{i=1}^{S} b_i s_i$$

$$P(R|Q,C) = \frac{e^{logO(R|Q,C)}}{1 + e^{logO(R|Q,C)}}$$

Formula 7. Logistic regression model[9].

b_0 is the intercept term, bi coefficients statistics and Si is the S series.

XIII. TREE MATCHING IN THE XML RETRIEVAL

Many problems should be examined with retrieval systems in relation to the problem of tree matching and structured search. Documents may be very large in size and when the search is not selective, the response may be composed of many results.

Xml document collection may include documents that are not specifically adapted to the structural search. Therefore, one of the key issues is how to choose the components that are approximately consistent with the limitations of search [35]. Tree matching algorithms are associated with the XML retrieval divided into two main sections. The first section covers the exact algorithms to find all patterns in a database XML. The second part describes and shows in detail approximation algorithm [36]. Branching pattern of the XML existing algorithms can be divided to the two-step algorithms of analysis approaches and one step algorithms of navigation approach. Evaluation of tree matching algorithms and approaches can be done in two ways for evaluating the performance and effectiveness. The exact tree matching approach is directly related to the efficiency while the approximate tree matching is more associated with effective [37].

XIV. CONCLUSION

At first, information retrieval was a matter for medical professionals, law, and library science. Users who worked less in secret, and more were seeking to study in the domain, were limited few via the companies of static documents and linguistic tools. But by appearence the era of information and expand use of the Internet, an abrupt mutation of the users number is developed, in general leading to the importance of discipline.

The World Wide Web and the Internet have brought to us a huge flood of data flow and aspects of life. So, unprecedented demand for efficient techniques to handle the enormous amounts of data is available. Nowadays querying the data and extracting relevant documents, is not enough. Users want to focus more on certain information even the smallest details that are irrelevant. XML that is seemed as semi-structured data, a potential candidate to meet these requirements.

This study is summarized on recent efforts in the field of XML retrieval. Also, the database community presents methods based on the use of traditional database techniques to XML data. Topics of interest include query languages such as SQL and referential data integrity problems. On the other hand, the IR community applies IR standard techniques with some variations to Centralized retrieval on the level of element. Despite some similarities with unstructured text, XML requires special attention, an aspect of determining relationship between the elements of the user's query and methods for its evaluation.

REFERENCES

[1] N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, Advances in XML Information Retrieval and Evaluation: 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Dagstuhl ... Papers (Lecture Notes in Computer Science): Springer-Verlag New York, Inc., 2006.

[2] R. van Zwol and T. van Loosbroek, "Effective Use of Semantic Structure in XML Retrieval," in Advances in Information Retrieval. vol. 4425, G. Amati, C. Carpineto, and G. Romano, Eds., ed: Springer Berlin Heidelberg ,2007 , pp. 621-628.

[3] M. S. Ali, M. P. Consens, and B. Helou, "Improving the Effectiveness of XML Retrieval with User Navigation Models," in Data Engineering, 2009. ICDE '09. IEEE 25th International Conference on, 2009, pp. 1584-1587.

[4] M. P. Consens ,G. Xin, Y. Kanza, and F. Rizzolo, "Self Managing Top-k (Summary, Keyword) Indexes in XML Retrieval," in Data Engineering Workshop, 2007 IEEE 23rd International Conference on, 2007, pp. 245-252.

[5] N. Naffakhi and R. Faiz, "Using Bayesian networks theory for aggregated search to XML retrieval," presented at the Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, Craiova, Romania, 2012.

[6] Z. Szlávik, A. Tombros, and M. Lalmas, "Investigating the use of summarisation for interactive XML retrieval," presented at the Proceedings of the 2006 ACM symposium on Applied computing, Dijon, France, 2006.

[7] A. Trotman and B. Sigurbjörnsson, "Narrowed Extended XPath I (NEXI)," in Advances in XML Information Retrieval. vol. 349 ,3N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, Eds., ed: Springer Berlin Heidelberg, 2005, pp. 16-40.

[8] A. Trotman and B. Sigurbjörnsson, "NEXI, Now and Next," in Advances in XML Information Retrieval. vol. 3493, N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, Eds., ed: Springer Berlin Heidelberg, 2005, pp. 41-53.

[9] S. Pal and M. Mitra, "XML Retrieval: A Survey," Citeseer2007.
[10] J. Pehcevski, J. Thom, S. M. M. Tahaghoghi, and A.-M. Vercoustre, "Hybrid XML Retrieval Revisited," in Advances in XML Information Retrieval. vol. 3493, N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, Eds., ed: Springer Berlin Heidelberg, 2005, pp. 153-167.
[11] L. M. de Campos, J. M. Fernández-Luna, J. F. Huete, and C. Martín-Dancausa, "Managing structured queries in probabilistic XML retrieval systems," Information Processing & Management, vol. 46, pp. 514-532, 9// 2010.
[12] W. Yanlong, L. Xinkun, C. Xiangrui, Z. Ying, and Y. Xiaojie, "XML Retrieval with Structural Context Relaxation," in Emerging Intelligent Data and Web Technologies (EIDWT), 2013 Fourth International Conference on, 2013, pp. 747-752.
[13] S. Amer-Yahia, N. Koudas, Am, l. Marian, D. Srivastava, et al., "Structure and content scoring for XML," presented at the Proceedings of the 31st international conference on Very large data bases, Trondheim, Norway, 2005.
[14] D. Carmel, Y. S. Maarek, M. Mandelbrod, Y. Mass, and A. Soffer, "Searching XML documents via XML fragments," presented at the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, Toronto, Canada, 2003.
[15] M. Salem, A. Woodley, and S. Geva, "IR of XML Documents – A Collective Ranking Strategy," in Advances in XML Information Retrieval. vol. 3493, N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, Eds., ed: Springer Berlin Heidelberg, 2005, pp. 113-126.
[16] A. Woodley and S. Geva, "ComRank: metasearch and automatic ranking of XML retrieval system," in Cyberworlds, 2005. International Conference on, 2005, pp. 8 pp.-154.
[17] J. A .Aslam and M. Montague, "Models for metasearch," presented at the Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, USA, 2001.
[18] G. Kazai, N. Govert, M. Lalmas, and N. Fuhr, "The INEX Evaluation Initiative," in Intelligent Search on XML Data. vol. 2818, H. Blanken, T. Grabs, H.-J. Schek, R. Schenkel, and G. Weikum, Eds., ed: Springer Berlin Heidelberg, 2003, pp. 279-293.
[19] M. S. Ali, M. Consens, X. Gu, Y .Kanza, F. Rizzolo, and R. Stasiu, "Efficient, Effective and Flexible XML Retrieval Using Summaries," in Comparative Evaluation of XML Information Retrieval Systems. vol. 4518, N. Fuhr, M. Lalmas, and A. Trotman, Eds., ed: Springer Berlin Heidelberg, 2007 ,pp. 89-103.
[20] Y. Mass and M. Mandelbrod, "Relevance Feedback for XML Retrieval," in Advances in XML Information Retrieval. vol. 3493, N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, Eds., ed: Springer Berlin Heidelberg, 2005, pp. 303-310.
[21] J. Kamps ,M. Marx, M. d. Rijke, r. Sigurbj, et al., "Structured queries in XML retrieval," presented at the Proceedings of the 14th ACM international conference on Information and knowledge management, Bremen, Germany, 2005.
[22] Y. Mass and M. Mandelbrod, "Component Ranking and Automatic Query Refinement for XML Retrieval," in Advances in XML Information Retrieval. vol. 3493, N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, Eds., ed: Springer Berlin Heidelberg, 2005, pp. 73-84.
[23] C. D. Manning, P. Raghavan, H. Sch and tze, Introduction to Information Retrieval: Cambridge University Press, 2008.
[24] S. Geva, "GPX – Gardens Point XML IR at INEX 2005," in Advances in XML Information Retrieval and Evaluation. vol. 3977, N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, Eds., ed: Springer Berlin Heidelberg, 2006, pp. 240-253.
[25] H. Tanioka, "A Fast Retrieval Algorithm for Large-Scale XML Data," in Focused Access to XML Documents, F. Norbert, K. Jaap, L. Mounia, and T. Andrew, Eds., ed: Springer-Verlag, 2008, pp. 129-137.
[26] P. Ogilvie and J. Callan, "Hierarchical Language Models for XML Component Retrieval," in Advances in XML Information Retrieval. vol. 3493, N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, Eds., ed: Springer Berlin Heidelberg, 2005 ,pp. 224-237.
[27] Y. Mass and M. Mandelbrod, "Using the INEX Environment as a Test Bed for Various User Models for XML Retrieval," in Advances in XML Information Retrieval and Evaluation. vol. 3977, N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, Eds., ed: Springer Berlin Heidelberg, 2006, pp. 187-195.
[28] J. Lafferty and C. Zhai, "Document language models, query models, and risk minimization for information retrieval," presented at the Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, USA, 2001.
[29] F. Huang, "Using Language Models and Topic Models for XML Retrieval," in Focused Access to XML Documents. vol. 4862, N. Fuhr, J. Kamps, M. Lalmas, and A. Trotman, Eds ,.ed: Springer Berlin Heidelberg, 2008, pp. 94-102.
[30] B. o. Sigurb jornsson, J. Kamps, and M. de Rijke, "University of Amsterdam at INEX 2005: AdHoc Track," in Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), 2006.
[31] C. Crouch, A. Mahajan, and A. Bellamkonda, "Flexible Retrieval Based on the Vector Space Model," in Advances in XML Information Retrieval. vol. 3493, N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik ,Eds., ed: Springer Berlin Heidelberg, 2005, pp. 292-302.
[32] T. Schlieder and H. Meuss, "Querying and ranking XML documents," J. Am. Soc. Inf. Sci. Technol., vol. 53, pp. 489-503, 2002.
[33] T. Wichaiwong and C. Jaruskulchai, "A simple approach to optimize XML Retrieval," in Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on, 2010, pp. 426-431.
[34] W. Lu, S. Robertson, and A. MacFarlane, "Field-Weighted XML Retrieval Based on BM25," in Advances in XML Information Retrieval and Evaluation. vol. 3977, N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, Eds., ed: Springer Berlin Heidelberg, 2006, pp. 161-171.
[35] M. A. Tahraoui, K. Pinel-Sauvagnat, C. Laitang, M. Boughanem, H. Kheddouci, and L. Ning, "A survey on tree matching and XML retrieval," Computer Science Review, vol. 8, pp. 1-23, 2013.
[36] C. Zhang, J. Naughton, D. DeWitt, Q. Luo, and G. Lohman, "On supporting containment queries in relational database management systems," presented at the Proceedings of the 2001 ACM SIGMOD international conference on Management of data, Santa Barbara, California, USA, 2001.
[37] S. Al-Khalifa, H. V. Jagadish, N. Koudas, J. M. Patel, D. Srivastava, and W. Yuqing, "Structural joins: a primitive for efficient XML query pattern matching," in Data Engineering, 2002. Proceedings. 18th International Conference on, 2002, pp. 141-152.