WHU-XML: an XML based Digital Library System

LU Wei，LIU Dan，FANG Fang，LONG Quan，YUAN Zelin，ZHANG Mi
Center for Studies of Information Resources, Wuhan University, Wuhan, 430072, China；
reedwhu@gmail.com; danliu.whu@gmail.com;darwin.fang@googlemail.com

Abstract

As a semi-structured data standard rich in both content and structure, XML is becoming the dominant information organization format in digital library. Compared with traditional information retrieval systems, XML retrieval system has great advantages in organizing and retrieving information due to its element-level rather than document-level access to relevant information. This paper gives a detailed introduction of an XML-based digital library system named WHU-XML, including the indexing of XML documents and image data, the retrieval of information and the presentation of search results.

1. Introduction

XML (eXtensible Markup Language) is a simple, but very flexible text format derived from SGML which originally designed to meet the challenges of large-scale electronic publishing [1]. It's now widely used in digital libraries, to describe digital library resources [2], to build a log format [3] and to represent data [4]. Also, it's used in information storage, information interchange, knowledge representation and other aspects in digital library.

In this paper, we propose an XML-based digital library system in which resources are XML formatted. Benefiting from this kind of storage method, our system provides element-level retrieval function, that is, user can get access to passages which he (she) will be most interested in. This function will undoubtedly facilitate the communication between user and documents.

The rest of this paper is organized as follows: Section 2 introduces the related work about XML-based digital library. Section 3 covers our system architecture. Specifically, we elucidate the indexing and retrieval method, as well as the human interface of our system in this section. Section 4 shows 2 graphs from the system implementation view and briefly introduced the evaluation we have done in some early papers. Finally, section 5 summarizes our proposed system and briefly introduces our future work.

978-1-4244-2511-2/08/$25.00 ©2008 IEEE

2. Related work

Recently, a number of XML-based digital library researches have been done. Ioannis Papadakis and Vassilios Chrissikopoulos in Department of Informatics of University of Piraeus proposed a framework for the development of XML-based digital libraries in order to "exploit the advantages that derive from the new XML standard in the field of digital libraries [5]". In addition, Jacky C.K. Ma and Michael R. Lyu in The Chinese University of Hong Kong designed and implemented an XML-Based Digital Video Library System which includes Video Server, Indexing Server, Query Server and Client Applications [6].

As a very important part of an XML-based digital library, XML retrieval has drawn much attention. Max-Planck Institute for Informatics developed a search engine for ranked retrieval of XML, named TopX, which "supports a probabilistic-IR scoring model for full-text content conditions and tag-term combinations, path conditions for all XPath axes as exact or relaxable constraints, and ontology-based relaxation of terms and tag names as similarity conditions for ranked retrieval" [7][8]. Hyper-media Retrieval Engine for XML (HyREX) is another XML retrieval system which is written in Perl and supports CO (Content Only) and CAS (Content And Structure) retrieval. The architecture of HyREX is quite similar to database management system and consists of HyGate, XIRQL, and HyPath [9]. Department of computer science in Cornell University developed the XRANK system which provides keyword search for XML documents. XRANK includes three components: ElemRank Computation, Hybird Dewey Inverted List and Query Evaluator [10].

3. System architecture

3.1. Document collection

This system uses two document collections provided by INEX [11] from 2002 to 2008: IEEE collection and Wikipedia collection.

The IEEE document collection is so far made up of the full-texts, marked up in XML, of 12,107 articles of the IEEE Computer Society's publications from 12 magazines and 6 transactions, covering the period of 1995-2002, and totaling 494 megabytes in size[12].

The Wikipedia collection is so far made up of the full-texts, marked up in XML, of 659,388 articles of the Wikipedia project, covering a hierarchy of 113,483 categories, and totaling more than 60 Gigabytes (4.6 Gigabytes without images). The collection has a structure containing text, more than 300,000 images and some structured parts corresponding to the Wikipedia templates (about 5000 different tags)[13].

3.2. Indexing

As a semi-structured data standard, XML is rich both in content and structure. In order to provide XML structural retrieval, the indexing of XML documents must be processed both on content and structure.

The indexing of content is just like the traditional indexing of plain texts. Inverted files are used to store the keywords and their occurrences in documents.

As to the indexing of structure of XML documents, the authors developed a B+ tree like, structured index from scratch. It trades the paths in XML documents as keywords, and uses inverted file to store the occurrence of these paths in the collection. Lu et al. [14][15] give a detailed introduction of the indexing methods, indexing procedure and the structure of the inverted files used in WHU-XML.

As to the images in Wikipedia collection, the system built two kinds of indexes. The first one listed the occurrence of images in the XML documents ordered by document ID. This was done when indexing content information by identifying the “figure” element, which was used to describe an image entity in the Wikipedia collection. We use the first kind of indexes to support text-based image retrieval. The second kind is color index of the images. An image can be viewed as a combination of different colors. To represent the feature of color, we used color histogram which is widely used in many systems. And we extracted 72-dimensional colors’ values to represent the image (more sophisticated image retrieval systems may use texture features, form features and spatial features to represent an image).

3.3. Retrieval

3.3.1. Text retrieval

(1) CO: Keyword search

The keyword search of WHU-XML system also called CO (Content Only) retrieval. Just as the traditional information search, user uses natural language to inquire. However, the returned units are no longer on document-level, but element-level, which means users can get access to the exact part they will be most interested in instead of browsing the whole article.

In the backward, the retrieval system took each XML element as a single document. When given a query, it will return the most relevant elements. However, there will be lots of overlapped elements as XML is nested. So, the system uses some choosing algorithm to get the most suitable element for the query.

The retrieval model used in WHU-XML is field-weighted BM25, which was proposed by Robertson et al. at INEX 2005. Lu et al. [16][17] give a detailed introduction about this model, and some improvements on the model in order to resolve some typical problems in element retrieval.

(2) CAS: Keyword search with structural constraints

The structural search of WHU-XML system is also called CAS (Content and Structure) retrieval. User can give structural constraints to keywords, following NEXI language [18] which is the official XML query language of INEX. For example the query “//article[., about(information retrieval)]//section[., about(XML)]” means to retrieve those kinds of document sections which are talking about “XML” and in an article talking about “information retrieval”.

There are two ways of interpretation for structural constraints in WHU-XML—vague interpretation and strict interpretation, as we are not sure whether the retrieved results are more satisfactory for a user when retrieving algorithms take the structural constrains into consideration. Vague interpretation means the system should retrieve keywords freely and then filter the results by judging whether the result element has the same structure as or is the child nodes of the structure the NEXI query specified. Strict interpretation means when looking for keywords, it should be done according to the structural constraints of that keyword, and the results should be in accordance with the NEXI query specified structure.

For example, given a NEXI query as follows:

"//article[.,about(information retrieval)]//section [.,about(XML)]"

When doing vague CAS retrieval, the system looks up "information", "retrieval", and "XML" respectively in the inverted file indexes and gets the elements where these three words show up. Then it calculates the score of each element according to field weighted BM25 Model and ranks these results according to their scores. Before returning the results to the user, the system does a filtering process in which the elements are judged whether they are the same or the child nodes of "//article//section". If not, discard the element. The remaining elements are returned to the user.

When doing strict CAS retrieval, the system looks up "information" in "//article", "retrieval" in "//article", and "XML" in "//article//section", then those elements which contain "XML" in section and also in a document talking about "information" and "retrieval" are picked out. The system then gives these elements scores and ranks them from high score to low score, and finally return them to the user.

3.3.2. Image retrieval

(1) TBIR (Text-Based image retrieval)

TBIR takes the text around the images as the description. TBIR systems first find the relevant information and then extract pictures around the relevant information. Two kinds of image retrieval approaches are allowed in WHU-XML: by specify the element level for retrieved image or by specify the relation between the retrieved relevant element and context.

In the first approach, users can select "body", "section" or "figure" as the image extraction level. "body", "section" and "figure" are element tags in the Wikipedia collection. For example, if a user wants to search images about "the great wall" and specifies the level of "body", the system firstly gets relevant elements about "the great wall" using the keyword search method described in 3.3.1, then gets all XML documents which these elements belong to, and only returns images in the "body" elements of these XML documents as search results. In this process, the score of a relevant element is regarded as scores of all images in the same document directly. If there is more than one relevant element in one document, the system will use a combination method to combine the scores.

In the second approach, relations such as self, sibling, parent, ancestor, 3rd ancestor etc. can be selected. For a relevant element in the result of keyword search, self means extract images from the element; sibling means extract images from its adjacent sibling elements; parent means extract images from the parent element and so on.

(2) CBIR (Content-Based image retrieval)

CBIR takes the images as a combination of colors. The query of CBIR is not keyword, but an image uploaded to the system by the user. Given a query image, the system extracts 72-dimensional colors' values and calculates the Euclidean distance between the given image and images in collection. The score of an image is the complement of the distance. That is, closer distance between two images means more similarity and higher score.

(3) TBIR+CBIR

The combinative retrieval of TBIR and CBIR uses a parameter to merge the retrieve results of the two methods. When user gives keywords, an image and the combinative parameter, the system first returns the result sets of TBIR and CBIR separately. Then, both result sets are normalized by dividing the highest score in the set. Finally, every image gets its own score calculated by a linear merge formula.

3.4. User interface

3.4.1. Text retrieval: About search, we gave a clear and tidy interface like most search engines, with some explanation of how to use the system below the search box (figure 1).

Figure 1. Main page of WHU-XML

About result browsing, we provided a bunch of browsing choices for the user to determine how to display the search results. These browsing choices are designed according to the sub tasks of ad-hoc retrieval in INEX.

(1) Overlap: without considering the overlap of elements, just ranking all the relevant elements according to their scores.

(2) Focused: non overlapped elements, and rank the elements from high score to low score.

(3) Browse Overlap: grouped the results of ① by the document ID. Documents are ranked according to article-level score, and in a document, relevant elements are ranked according to element-level score.

(4) Browse Focused: grouped the results of ② by the document ID.

(4) Best in context: there is only one element returned for one document, and this element is the best point for a user to begin reading the article.

Besides different browsing choices, WHU-XML system differ traditional retrieval systems also in that the returned results are on element level rather than document level. When reading the information, users can choose to just read the relevant element part, the whole document or even the XML document (figure 2).

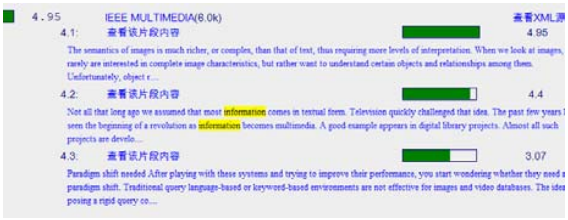

Figure 2. Browse focused

3.4.2. Image retrieval: Our system offers two approaches, searching by level or by relation, for users to perform their image search, as mentioned in 3.2.2. If the user wants to retrieve some images by the CBIR way, he (she) can upload the image to the system or select one image from the image set directly, then the system will give back a bunch of images have the similar colors.

For image retrieval, we give a browsing way just like Google Image. Additionally, users can browse metadata information of an image, and the XML element which contains that image (figure 3).

Figure 3. Image retrieval result

4. System implementation and evaluating

4.1. System implementation

WHU-XML was developed by using Java and C, that is, the front interface was written in Java and the efficient indexing and retrieving are programmed using C under Linux. Figure 4 shows the architecture of the text retrieval part and figure 5 shows the architecture of the image retrieval part.

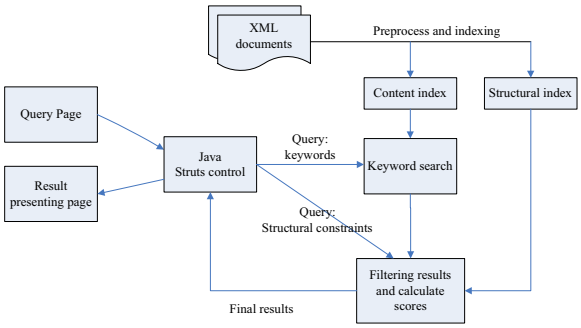

Figure 4. Text retrieval system architecture

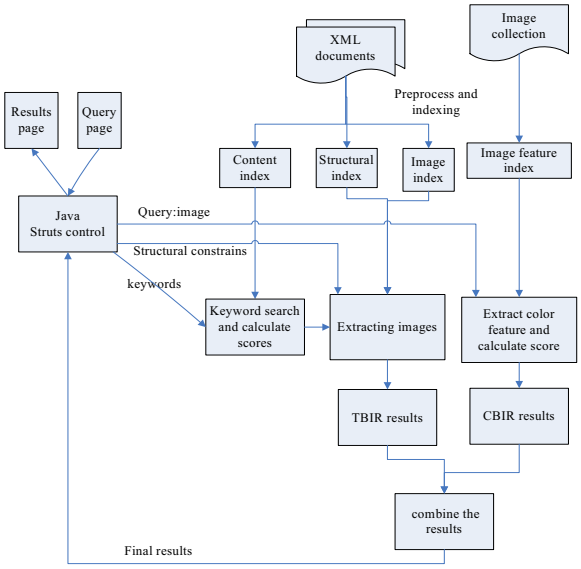

Figure 5. Image retrieval system architecture

4.2. Evaluating

We did experiments from 3 aspects respectively: the indexing speed and size, the effectiveness of text retrieval and the effectiveness of image retrieval. In [15] we described the indexing time and size of Shakespeare's works and the IEEE datasets described in 3.1. In [16] and [19], the topics of INEX 2004[12] and INEX 2005[20] are evaluated by different measures including AveP[21], nxCG and ep-gr[22] for

the evaluation of text retrieval effectiveness. In [23] we discussed about the effectiveness of image retrieval by using 20 topics measured by Precision, MAP and P@10.

5. Conclusion and future work

As an XML-based digital library system, WHU-XML provides various retrieving functions and plenty of browsing ways, and it has a lot of advantages compared with traditional retrieval systems both in the organization and retrieval of information.

However, there are still lots of work to do in the future. Firstly, we may extend the data collection to Chinese environment or different XML structures, such as academic papers, dissertations and so on. Secondly, as a system not only for theorical research but also for practical use, speed, efficiency and user interface need more work. Thirdly, the image retrieval part can be expanded by adding texture feature and form feature of images to the image retrieval model.

6. References

[1] Extensible Markup Language: http://www.w3.org/XML/.

[2] Chadia Moghrabi, "Digital Library Resources Description", Proceedings of the International Conference on Information Technology, Coding and Computing (ITCC'04), The Orleans, Las Vegas, Nevada, USA., 2004.

[3] Marcos G, Ganesh P, Filip J., and Lillian C, "The XML Log Standard for Digital Libraries", Analysis, Evolution, and Deployment. Proceedings of the 2003 Joint Conference on Digital Libraries (JCDL'03), Houston, Texas, USA, 2003.

[4] Abdel B, Ingrid F and Yves R, "XML Data Representation in Document Image Analysis", Proceedings of Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Parana, Brazil, 2007.

[5] Ioannis Papadakis and Vassilios Chrissikopoulos, "A Digital Library Framework based on XML". http://citeseer.ist.psu.edu/543571.html.

[6] Jacky C.K. Ma and Michael R. Lyu, "Design and Implementation of XML-Based Digital Video Library System", International Conference on Internet Computing (1) 2001.

[7] TopX Introduction: http://topx.sourceforge.net/.

[8] Theobald M, Schlemiel R and Weikum G, "TopX and XXL at INEX 2005", Proceedings of INEX2005, 2006, 282-295.

[9] Fuhr N, Govert N and Grobjohann K, "HyRex: Hyper-media Retrieval Engine for XML", SIGIR 2002, 2002, 449-449.

[10] Lin Guo, Feng Shao, Chavdar Botev and Jayavel Shanmugasundaram, "XRANK: Ranked Keyword Search over XML Documents", Proceedings of the 2003 ACM SIGMOD international conference.

[11] INEX 2008: http://www.inex.otago.ac.nz/.

[12] INEX 2004: http://inex.is.informatik.uni-duisburg.de:2004/.

[13] L. Denoyer and P. Gallinari, "The Wikipedia XML Corpus", SIGIR Forum 2006, 40:64–69.

[14] Lu Wei and Xia Lixin, "Implementation of OKAPI based XML Information Retrieval", The Journal of The Library Science In China, 2006,25(6):679-685.

[15] Lu Wei, "XML Indexing Based on Traditional IR System", Journal of information, 2006,25(6):679-685.

[16] W. Lu, S. Robertson and A. Macfarlane, "Field-Weighted XML Retrieval Based on BM25", Proceedings of INEX 2005, LNCS, 2006: 126-137.

[17] Lu Wei, "The Key Problems and Corresponding Solutions for XML Element Retrieval", The Journal of The Library Science In China, 2007(6).

[18] Andrew Trotman, and Börkur Sigurbjörnsson. "Narrowed Extended XPath I (NEXI)", Proceedings of the INEX 2004, (pp.16-40).

[19] Lu Wei, "Field-weighted XML Document Level Retrieval and Evaluation", The Journal of The Library Science In China, 2006(6).

[20] INEX 2005: http://inex.is.informatik.uni-duisburg.de/2005/.

[21] AP de Vries, G Kazai and M Lalmas, "Evaluation metrics 2004", Proceedings of the 3rd INEX Workshop, LNCS, pp.250-251.

[22] Gabriella Kazai and Mounia Lalmas, "INEX 2005 Evaluation Metrics", Fourth Initiative on the Evaluation of XML Retrieval (INEX), 2006.

[23] Lu Wei, Zhang Mi and Liu Dan. "The implementation and evaluation of image retrieval based on XML Fragment", (to appear).