Keyword-based XML Fragment Retrieval: Experimental Evaluation based on INEX 2003 Relevance Assessments

Kenji Hatano†, Hiroko Kinutani‡, Masahiro Watanabe§
Yasuhiro Mori♮, Masatoshi Yoshikawa♮, Shunsuke Uemura†

† Nara Institute of Science and Technology, Japan
{hatano,uemura}@is.aist-nara.ac.jp
‡ Japan Science and Technology Agency, Japan
kinutani@dblab.is.ocha.ac.jp
§ The National Institute of Special Education, Japan, masahiro@nise.go.jp
♮ Nagoya University, Japan
mori@dl.itc.nagoya-u.ac.jp, yosikawa@itc.nagoya-u.ac.jp

ABSTRACT

We have developed a keyword-based XML fragment retrieval system based on statistics of XML documents to improve both the efficiency and effectiveness of the system. Currently, relevance assessments for keyword-based XML fragment retrieval systems are provided only by the INEX project; thus we evaluate our system using them. However, our system performs poorly with respect to retrieval accuracy using the INEX 2003 relevance assessments. In this paper, we analyze all CO topics based on statistics of answer XML fragments and report our experimental results. After performing the experiments, we found that two types of CO topics is present in the INEX 2003 relevance assessments and has to be handled when in the experimental evaluations.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Information retrieval, Performance evaluation

Keywords

Keyword-based XML fragment retrieval, Evaluation of both efficiency and effectiveness, Analysis of relevance assessments

1. INTRODUCTION

Extensible Markup Language (XML) [5] is becoming widely used as a standard document format in many application domains. In the near future, we believe that a great variety of documents will be produced in XML; therefore, in a similar way to developing Web search engines, XML information retrieval systems will become very important tools for users wishing to explore XML documents on the Internet or a company intranet.

XQuery [4], proposed by the World Wide Web Consortium (W3C), is known as a standard query language for retrieving fragments of XML documents. Using XQuery, users can issue a flexible query consisting of both some keywords and XPath notations[1]. If users already have knowledge of the structure of XML documents, users can issue XQuery-style queries. However, there are a lot of XML documents whose XML schemas are different to each other, and as a result, nobody can issue such a formulated query into information retrieval systems. Consequently, we believe that XML retrieval systems should employ a much simpler form of query such as keyword search services. Keyword search services enable users to retrieve needed information by providing them with a simple interface. It is, therefore, the most popular information retrieval method, since users need to know neither a query language nor the structure of XML.

Because of the aforementioned background on XML information retrieval, close attention has recently been paid to a keyword-based XML fragment retrieval system. Some keyword-based XML fragment retrieval systems have already been made available. They assume the existence of the document type declaration, which contains or points to markup declarations that provide a document-type definition (DTD) for a class of XML documents. As a result, they can deal with only one type of XML documents. It is true that the DTD facilitates the enhancement of retrieval accuracy and retrieval speed of keyword-based XML fragment retrieval systems. However, XML documents on the Internet or a company intranet do not always include DTDs; thus they cannot deal with multiple types of XML documents whose structures are different to each other. Because the XML documents feature many types of document structures, a next-generation Web search engine will have to treat XML documents whose structures are different.

To cope with the problems described above, we have developed a keyword-based XML fragment retrieval system using statistics of XML documents [12]. In XML fragment retrieval, we assume that users explicitly specify query keywords; thus, we believe that the size of retrieval results become small compared to document retrieval. On the other hand, we also believe that extremely small XML fragments have neither rhyme nor reason by themselves. To solve this

[1] Currently, the XML Query working group is just starting to develop full-text search functions [2, 6].

problem, we designed an XML fragment retrieval system that can return small (but not extremely small) and semantically useful XML fragments as retrieval results.

According to [14], the INEX 2002 relevance assessments tended to regard large-size XML fragments as correct retrieval results. This fact did not meet purpose of our XML fragment retrieval system. If the INEX 2003 relevance assessments are also similar to the previous one, our system might perform poorly in its retrieval accuracy. It is therefore necessary to analyze and evaluate CO topics of the INEX 2003 relevance assessments to determine whether they suit our purpose.

In this paper, we analyze the INEX 2003 relevance assessments based on their statistics, and evaluate our system using CO topics reflecting our analyses. We believe that analyzing the relevance assessments helps to both construct the next version of the relevance assessments and improve the efficiency and effectiveness of our system.

The remainder of this paper is organized as follows. First, we describe our keyword-based XML fragment retrieval system in Section 2. Then, we report analyses of the INEX 2003 relevance assessments in Section 3 and discuss controversial points of CO topics of the INEX 2003 relevance assessments in Section 4. Finally, we conclude this paper in Section 5.

2. OUR XML FRAGMENT RETRIEVAL SYSTEM

This section introduces proposed retrieval model and the purpose of our keyword-based XML fragment retrieval system. We also report preliminary experimental results obtained using our system, and explain our observations regarding XML fragment retrieval.

2.1 Data Model and Retrieval Model

For simplicity, our system's data model is similar to that of the XPath data model [7], because XML is modeled as a hierarchical tree. Actually, the only difference between the XPath data model and ours is that attribute node is regarded as a child of an element node[2].

In addition, for the sake of easy comprehension, our system's retrieval model bears a close resemblance to the proximal nodes model [17]. Basically, our logical model of an XML fragment is a sub-tree whose root node is an element node. Our system can identify XML fragments by their reference numbers derived from document order; therefore, our system can obtain similarities between user's query and XML fragments based on their document orders.

2.2 Purpose of Our Research

We distinguish two types of keyword-based XML fragment retrieval systems. This paper refers to these two types of systems as *XML data retrieval systems* and *XML information retrieval systems*, for the sake of convenience. The former is based on structured or semi-structured database systems with keyword proximity search functions that are modeled as labeled graphs, where the edges correspond to the relationship between an element and a sub-element and to `IDREF` pointers [1, 11, 13]. Dealing with XML documents as XML graphs facilitates the development of keyword-based information retrieval systems, which are able to perform the retrieval processing efficiently. On the other hand, the latter has been developed in the research field of information retrieval [8, 9], and enables us to retrieve XML fragments without indicating element names of XML documents. The large difference between the XML retrieval systems and the XML information retrieval systems derives from data characteristics of their retrieval targets. In short, we consider that the former focuses mainly on *data-centric* XML documents, whereas the latter deals with *document-centric* ones[3]. In the meanwhile, almost all XML data retrieval systems and XML information retrieval systems currently assume the existence of DTD of XML documents. It is a fact that DTD facilitates enhancing retrieval accuracy and retrieval speed of their systems. However, there are some problems associated with searching XML fragments on the Internet or a campany intranet, as described in Section 1; thus other types of XML retrieval systems, which do not utilize DTD, are required. Consequently, XML retrieval systems in the future will have to deal with XML documents whose structures are different.

To meet the needs of the new architecture of XML retrieval systems, we have developed a keyword-based XML fragment retrieval system using statistics of XML documents [12]. Our system focuses on retrieval of document-centric XML documents rather than that of data-centric ones, and does not utilize any information in relation to element names of XML documents, whereas the systems introduced above take advantage of such information for querying and indexing of XML documents. Our approach dictates that XML documents must be divided into fragments in order to develop a keyword-based XML retrieval system. Because XML is a markup language, XML documents can be automatically divided into their fragments using their markup [15]; a problem surfaces, however, because this gives rise to an unmanageable profusion of XML fragments. In other words, it takes very long time to retrieve XML fragments related to a keyword-based query using our approach. For this reason, not inspecting extracted all XML fragments, but retrieving only XML fragments which are informative enough for XML information retrieval would be better.

2.3 Evaluating Our System based on INEX 2003 Relevance Assessments

In this section, we report the retrieval accuracy of our keyword-based XML fragment retrieval system based on INEX 2003 relevance assessments. The relevance assessments defined two metrics, strict and generalized; thus we performed experimental evaluations based on both metrics. The metrics have two criteria, "exhaustiveness" and "specificity," for quality metrics of IR applications. The method of how recall and precision are computed is described in a report [18][4]. Based on the metrics, we drew recall-precision curves

[2]If element node has some attribute nodes that have brotherhood ties, they are owned by the element node.

[3]There is a data-centric and a document-centric view of XML described in [3].

[4]Another way is also available, described in a technical report [10]; however, we did not apply it in this paper.

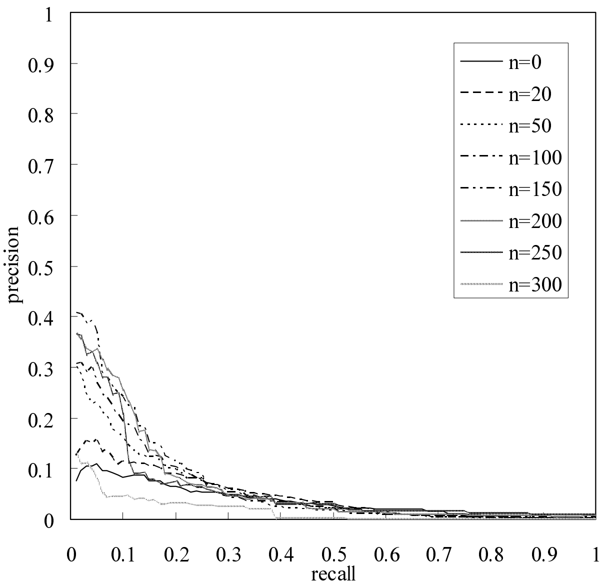

Figure 1: Evaluation of our system based on INEX 2003 relevance assessments (strict).

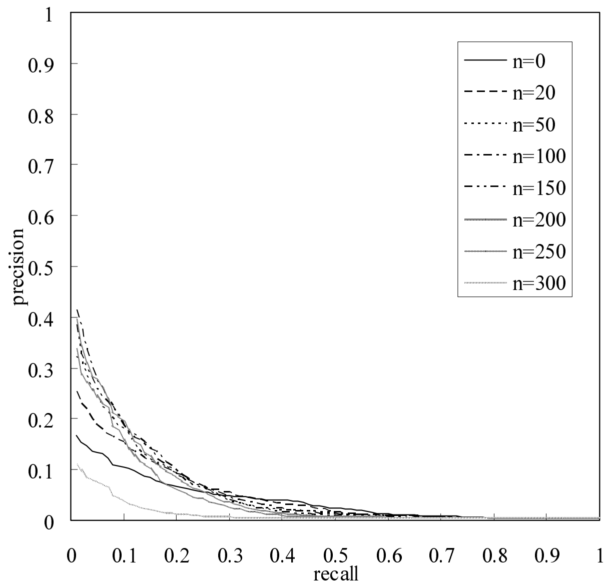

Figure 2: Evaluation of our system based on INEX 2003 relevance assessments (generalized).

to evaluate the XML fragment retrieval system. Figure 1 and 2 show recall-precision curves of our system based on INEX 2003 relevance assessments. In these figures, n means the minimum number of tokens, which is defined to eliminate extremely small XML fragments from retrieval targets[5]. In short, as n becomes larger, the retrieval accuracy of our system also increases, as does our system's retrieval speed.

As shown in these two figures, we think that our keyword-based XML fragment retrieval system may be performing poorly. Although we recognize the problems inherent in our system[6], it is thought that the problems may not only reside in our system, but also in the relevance assessments. If the INEX 2003 relevance assessments tend to regard large-size XML fragments as correct retrieval results in analogy with the INEX 2002 relevance assessments, our system will be a poorly-performing XML fragment retrieval system, because our system tends to retrieve small XML fragments, but not extremely small ones, as retrieval results described in Section 1. As a matter of fact, in the case where the threshold of the number of tokens is 150 (in strict relevance assessments) or 100 (in generalized ones), our system does work properly (see Table 1). From our perspective, we consider that this number of tokens is very large for XML fragment retrieval, because the number of tokens is comparable to XML fragments whose root node is `ss1`, `ss2`, or `ss3`, as shown in Table 2. We have designed our XML fragment retrieval to enable users to retrieve XML fragments corresponding to other XML fragments whose size is less than (sub)sections of the INEX document collection as retrieval results. Therefore, we can forecast that retrieval results by our system would give fewer XML fragments than answer XML fragments determined in the INEX 2003 relevance assessments[7].

Table 1: Average precision of our system.

n	strict	generalized
0	0.0356	0.0390
20	0.0436	0.0476
50	0.0502	0.0505
100	0.0568	<u>0.0568</u>
150	<u>0.0669</u>	0.0525
200	0.0630	0.0503
250	0.0572	0.0416
300	0.0163	0.0130

Based on the aforementioned points, we analyze answer XML fragments of the topics of INEX 2003 relevance assessments, choose the topics suitable for our system, and re-evaluate our system's retrieval accuracy in reference to a revised version of the relevance assessments.

3. ANALYSES OF INEX RELEVANCE ASSESSMENTS

3.1 Analyses of the Relevance Assessments

As we described in the previous section, we consider that the INEX 2003 relevance assessments may work against XML fragment retrieval systems, which tend to regard small-size XML fragments as correct retrieval results. Consequently, we analyze answer XML fragments defined in the relevance assessments. Our system can deal with only content-only

[5]The size of XML fragments is proportional to the number of tokens contained in the XML fragments.

[6]Our system cannot calculate similarities between a query and XML fragments using only contents of XML documents.

[7]Of course, this is our opinion. In [16], the authors claimed that 500 words is valid for answer XML fragments. The proper size of answer XML fragments depends on the retrieval purposes of each INEX participant.

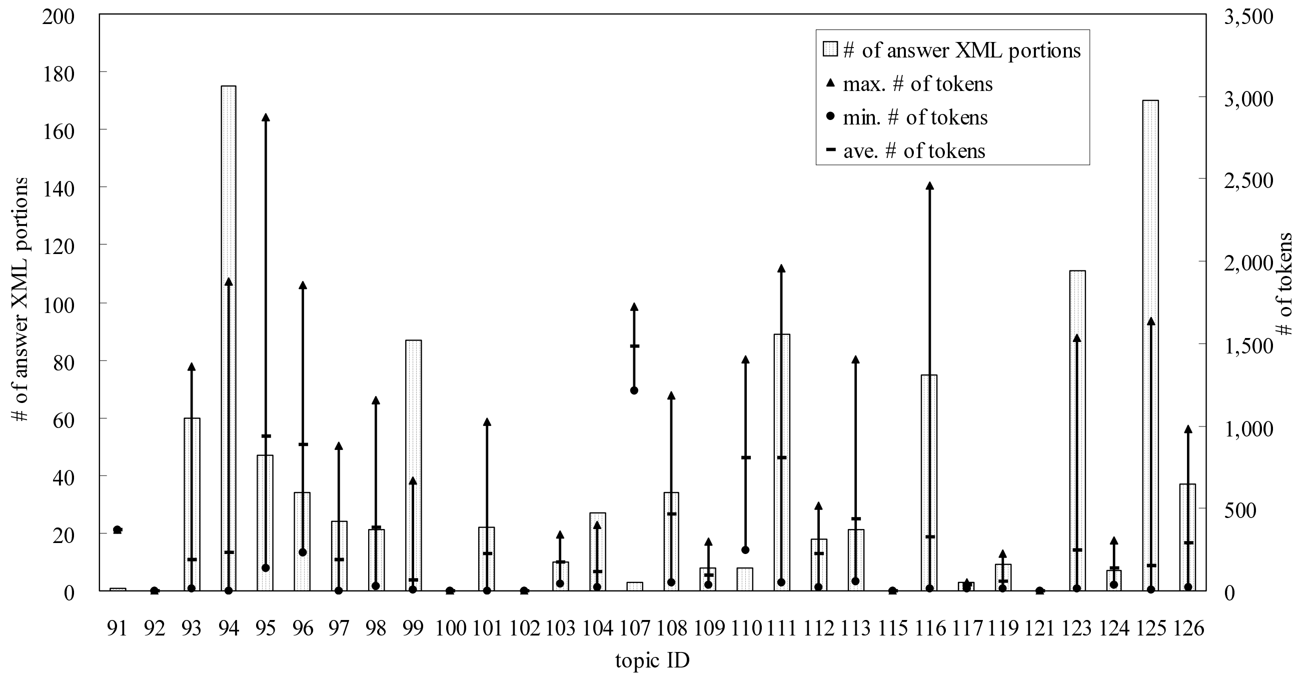

Figure 3: Analyses of the INEX 2003 relevance assessments.

(CO) topics of the relevance assessments; thus, only answer XML fragments of CO topics are analyzed here. In this section, we analyze the answer XML fragments whose exhaustiveness and specificity are 3.

Figure 3 shows analyses of CO topics of the INEX 2003 relevance assessments. Area bars represent the maximum, average, and minimum number of tokens of answer XML fragments, and line bars indicate the numbers of answer XML fragments. As shown in this figure, we first found that five CO topics of the relevance assessments (whose topic IDs are #92, #100, #102, #115, and #121) did not have answer XML fragments. It is doubtful that these topics were adopted as the CO topics of the relevance assessments, because we think that the CO topics with few answer XML fragments are not inappropriate for the relevance assessments. Moreover, we also found that the average number of tokens of almost all CO topics exceeded 100. In particular, the average number of tokens of CO topics whose IDs are #95, #96, #107, #110, and #111 was 500 and above; therefore, these CO topics distinctly work against our system. Furthermore, the number of answer XML fragments substantially differs with each CO topic.

From the aforementioned points, we select 14 CO topics (#93, #94, #97, #98, #99, #101, #104, #108, #112, #113, #116, #123, #125, #126) as the relevance assessments suitable for our system, and re-evaluate retrieval accuracy of our system based on them in the next section.

3.2 Reevaluation of Our System

Figures 4 and 5 show recall-precision curves of our system based on revised versions of INEX 2003 relevance assessments, while Table 3 shows average precision of each recall-precision curve. In comparison with previous evaluations described in Section 2.3, the retrieval accuracy of our system shows an improvement of 3.55% in this experiment. It is clear that our system performs better because of using selected CO topics. The thing which we want to assert is not improving retrieval accuracy of our system, but presence of two types of CO topics in XML fragment retrieval. In short, the CO topics for searching large-size XML fragments make our system worse retrieval accuracy of our system, which explains why our system could not return the XML fragments relevant to CO topics on the relevance assessments in the previous evaluations. Needless to say, we do not know whether our system's retrieval accuracy is better than that of other INEX participants' systems, though we could confirm controversial points of the relevance assessments for our system.

To reduce the scope of such arguments, we have to clarify what XML fragment retrieval is. It is difficult to define the granularity of XML documents for XML fragment retrieval; however, we think that it is important for INEX participants to determine topics of the relevance assessments suitable for their retrieval purposes and to automatically select the topics for evaluation by their respective retrieval systems. In the case of our keyword-based XML fragment retrieval system, small and semantically useful XML fragments are defined as correct retrieval results; thus, we consider that our only option is to use our relevance assessments explained in Section 3.1.

4. DISCUSSION

As we described in the previous section, retrieval accuracy of XML fragment retrieval systems depends on retrieval pur-

Table 2: Statistical analysis of XML fragments.

element	# of XML fragments	# of tokens		
		average	maximum	minimum
book	3,612,202	28,897	64,181	6,341
journal	6,314,623	7,342	14,903	3,982
article	11,801,575	974	4,727	29
bdy	9,271,423	765	3,943	11
index	72,993	623	1,593	230
bm	3,125,254	310	2,863	2
dialog	41,317	212	906	19
sec	14,078,415	201	2,613	1
bib	1,662,190	194	1,959	8
bibl	1,662,640	194	1,959	8
app	812,923	138	1,353	2
ss1	7,854,413	127	2,109	1
ss2	1,509,337	92	1,261	1
ss3	11,642	91	325	9
fm	797,123	65	289	9
tgroup	363,102	62	401	2
proof	229,144	60	801	5
vt	1,021,500	55	235	2
dl	18,670	52	745	5
edintro	28,923	50	272	4

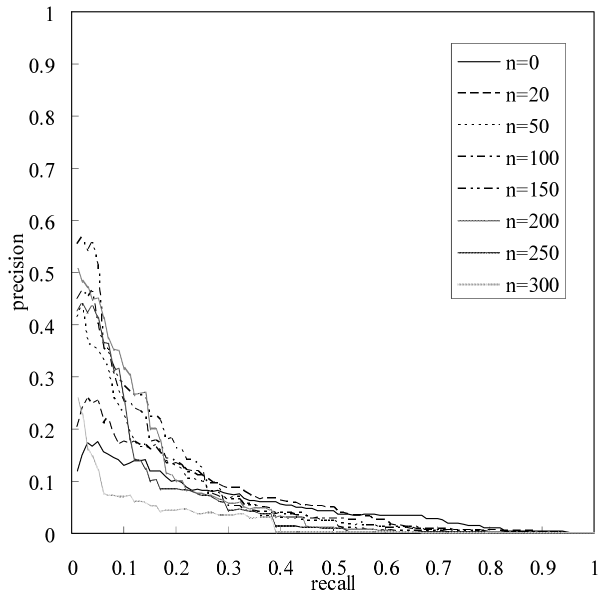

Figure 4: Evaluation of our system based on revised INEX 2003 relevance assessments (strict).

Table 3: Average precision of our system based on revised INEX 2003 relevance assessments.

n	strict	generalized
0	0.0564	0.0630
20	0.0666	0.0697
50	0.0689	0.0667
100	0.0777	<u>0.0774</u>
150	<u>0.0866</u>	0.0731
200	0.0769	0.0695
250	0.0611	0.0645
300	0.0253	0.0217

pose of each topic in the relevance assessments. In short, an XML fragment retrieval system performs poorly in retrieval accuracy if the retrieval purpose of an XML fragment retrieval system does not correspond to answer XML fragments of the topics. In the case of our system, small (not extremely small) and semantically useful XML fragments are retrieved as retrieval results. However, some topics tend to retrieve large XML fragments, meaning that our system performed poorly in retrieval accuracy. Therefore, we consider that the XML fragment retrieval systems that tend to regard large XML fragments as retrieval results gain the upper hand in retrieval accuracy. In this section, we make specific mention of the controversial points about CO topics of the INEX 2003 relevance assessments.

4.1 Characteristics of CO Topics

As we described in Section 3.1, the size and the number of answer XML fragments of CO topics vary (see Figure 3), though we do notice there are two types of CO topics in the relevance assessments (see Section 3.2). One is for searching specific XML fragments (SCO) and the other is for searching aggregated XML fragments (ACO). Query keywords of SCO topics may consist of some proper nouns, such as "Charles Babbage," "XML," and "Markov." Consequently, SCO topics tend to retrieve small-size XML fragments. On the other hand, ACO topics tend to retrieve exhaustive XML fragments, resulting in large answer XML fragments.

We think the existence of two types of CO topics are supported by the following statistical observations:

- Table 4 shows the topics of the INEX 2003 relevance assessments that have less than an average of 500 tokens in answer XML fragments whose $(exhaustiveness(E), specificity(S))$ is equal to $(3, 3)$. We consider that the sizes of the answer XML fragments are smaller than those of XML fragments whose $(E, S) = (3, 2)$ or $(3, 1)$, which means the average number of tokens of answer XML fragments should also be smaller.
- Table 5 shows the topics of the INEX 2003 relevance assessments that have more than an average of 500 tokens in the answer XML fragments. A search of these topics is exhaustive because they do not contain any specific keywords. That is to say, the answer XML fragments should cover information on the contents of the topic. As a result, we expect that the XML fragments, which are assessed as $(E, S) = (3, 3)$, become aggregated XML fragments with comparatively large granularity.

Current relevance assessments do not consider these controversial points, but we believe that considering these points helps to construct a high-quality test collection for XML fragment retrieval. If indeed a high-quality test collection can be constructed, XML fragment retrieval systems only have to deal with two such types of CO topics.

Table 4: SCO topics of the INEX 2003 relevance assessments.

topic ID	title	# of tokens (average)		
		(E, S)		
		(3, 3)	(3,2), (3, 1)	(2, 3)
93	"Charles Babbage" -institute -inst	186	3,377	62
94	"hyperlink analysis" +"topic distillation"	232	83	333
97	Converting Fortran source code	186	753	27
98	"Information Exchange" +XML "Information Integration"	383	0	347
99	perl features	69	314	18
101	+"t test" +information	228	364	222
104	Toy Story	114	735	0
108	ontology ontologies overview "how to" practical example	466	872	367
112	+"Cascading Style Sheets" -"Content Scrambling System"	228	332	61
113	"Markov models" "user behavior"	438	1,010	90
116	"computer assisted art" "computer generated art"	330	702	207
123	multidimensional index "nearest neighbor search"	245	546	48
125	+wearable ubiquitous mobile computing devices	154	249	47
126	Open standards for digital video in distance learning	288	710	455

Table 5: ACO topics of the INEX 2003 relevance assessments.

topic ID	title	# of tokens (average)		
		(E, S)		
		(3, 3)	(3,2), (3, 1)	(2, 3)
95	+face recognition approach	940	593	486
96	+"software cost estimation"	885	1,174	537
107	"artificial intelligence" AI practical application industry "real world"	1,487	0	633
110	"stream delivery" "stream synchronization" audio video streaming applications	811	669	162
111	"natural language processing" -"programming language" -"modeling language" +"human language"	806	474	253

4.2 Consistent Criteria

As we described in Section 2.3, there are two criteria, exhaustiveness and specificity, in the INEX 2003 relevance assessments. Both exhaustiveness and specificity have four levels, though the definitions of each level are too vague for us to accurately evaluate which XML fragments are relevant to a given topic.

For example, Table 6 shows the XML fragments that were assessed as $(E, S) = (3, 3)$ in topic #125. We found in Table 6 that there are some nested relationships among the XML fragments, though at that time, we did not understand how to allocate marks to the XML fragments related to the topic. We think that XML fragments related to a topic depend on retrieval purpose of the topic; therefore there are a lot of possible interpretation to what constitutes a good XML retrieval unit. Such interpretations cause confusion, thus preventing strict evaluations of XML fragment retrieval systems. In the INEX 2002 relevance assessments, two criteria, "relevance" and "coverage," feature rules between a parent node and its children nodes in the XML fragments; thus, we think the above problems did not occur. Consequently, the INEX 2003 relevance assessments also require standards of judgment with regard to exhaustiveness and specificity. At minimum, we consider that a definition is needed for determining which XML fragment is the most relevant to topic #125 in Table 6. That is to say, we believe that we should clearly define what the answer XML fragments is in the relevance assessments and should obtain consensus about it among INEX participants.

5. CONCLUSION

In this paper, we analyzed the INEX 2003 relevance assessments based on statistics of their answer XML fragments on CO topics, and reported some controversial points of the relevance assessments.

We strongly recommend judging the validity of each topic; in particular a topic that has no answer XML fragments at all or has only a few answer XML fragments is inadequate for the topics in relevance assessments. We also found that there are two types of CO topics for analyzing the relevance assessments. Our system tends to regard small XML fragments as retrieval results, thus it performs poorly in retrieval accuracy using the topics for searching aggregated XML fragments. Retrieval purposes of XML fragment retrieval systems are different, making it difficult to construct relevance assessments that meet the requirements of all XML fragment retrieval systems. However, we will be able to evaluate XML fragment retrieval systems if each system can choose the topics that fulfill its retrieval purpose. It is, therefore, necessary in constructing the INEX 2004 relevance assessments to define SCO and ACO topics whose numbers are the same. Of course, XML fragment retrieval systems should automatically judge types of topics and choose one to suit retrieval purpose of the topics by themselvs.

Moreover, we think that it is important to refine the INEX test collection year by year, requiring that excellent topics be selected from the INEX 2002/2003 relevance assessments, and be reused in INEX 2004. In conclusion, we need to define the baseline of excellent topics, in addition to adopting

Table 6: XML fragments evaluated $(E, S) = (3, 3)$ of topic #125.

file	path	# of tokens
co/1999/r1057	/article[1]	1128
co/1999/r1057	/article[1]/bdy[1]	863
co/1999/r1057	/article[1]/bdy[1]/sec[3]	215
co/1999/r1057	/article[1]/bdy[1]/sec[3]/fig[1]	37
co/1999/r1057	/article[1]/bdy[1]/sec[3]/fig[1]/art[1]	11
co/1999/r1057	/article[1]/bdy[1]/sec[3]/fig[1]/fgc[1]	24
co/1999/r1057	/article[1]/bdy[1]/sec[3]/p[1]	63
co/1999/r1057	/article[1]/bdy[1]/sec[3]/p[2]	49
co/1999/r1057	/article[1]/bdy[1]/sec[3]/p[3]	32
co/1999/r1057	/article[1]/bdy[1]/sec[3]/p[4]	33
co/1999/r1057	/article[1]/bdy[1]/sec[3]/p[5]	82
co/1999/r1057	/article[1]/bdy[1]/sec[5]	343
co/1999/r1057	/article[1]/bdy[1]/sec[6]	308
co/1999/r1057	/article[1]/bm[1]/app[1]/p[1]	65
co/1999/r1057	/article[1]/bm[1]/app[1]/p[2]	68
co/1999/r1057	/article[1]/bm[1]/app[3]/p[1]	34
co/1999/r1057	/article[1]/bm[1]/app[3]/p[2]	54
co/1999/r1057	/article[1]/bm[1]/app[3]/p[3]	25

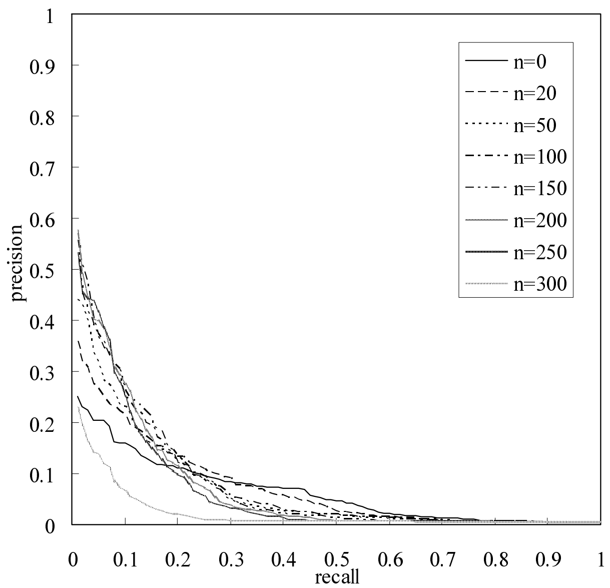

Figure 5: Evaluation of our system based on revised INEX 2003 relevance assessments (generalized).

the topics of INEX 2002/2003 relevance assessments that meet the baseline as topics of the INEX 2004 ones. Utilizing statistics of answer XML fragments is one solution.

6. ACKNOWLEDGMENTS

This work was partly supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, under grants #14780130, #14780325 and #15017243, and also supported by the Core Research for Evolutionary Science and Technology (CREST) Program "Advanced Media Technology for Everyday Living" of the Japan Science and Technology Agency (JST), Japan.

7. REFERENCES

[1] S. Agrawal, S. Chaudhuriand, and G. Das. DBXplorer: A System for Keyword-Based Search over Relational Databases. In *Proc. of the 18th International Conference on Data Engineering*, pages 5–16. IEEE CS Press, Feb./Mar. 2002.

[2] S. Amer-Yahia and P. Case. XQuery and XPath Full-Text Use Cases. `http://www.w3.org/TR/xmlquery-full-text-use-cases/`, Feb. 2003. W3C Working Draft 14 February 2003.

[3] H. M. Blanken, T. Grabs, H.-J. Schek, R. Schenkel, and G. Weikum, editors. *Intelligent Search on XML Data, Applications, Languages, Models, Implementations, and Benchmarks*, volume 2818 of *LNCS*. Springer-Verlag, Sep. 2003.

[4] S. Boag, D. Chamberlin, M. F. Fernandez, D. Florescu, J. Robie, and J. Siméon. XQuery 1.0: An XML Query Language. `http://www.w3.org/TR/xquery`, Nov. 2003. W3C Working Draft 12 November 2003.

[5] T. Bray, J. Paoli, C. M. Sperberg-McQueen, and E. Maler. Extensible Markup Language (XML) 1.0 (Second Edition). `http://www.w3.org/TR/REC-xml`, Oct. 2000. W3C Recommendation 6 October 2000.

[6] S. Buxton and M. Rys. XQuery and XPath Full-Text Requirements. `http://www.w3.org/TR/xquery-full-text-requirements/`, May 2003. W3C Working Draft 02 May 2003.

[7] J. Clark and S. DeRose. XML Path Language (XPath) Version 1.0. `http://www.w3.org/TR/xpath`, Nov. 1999. W3C Recommendation 16 November 1999.

[8] S. Cohen, J. Mamou, Y. Kanza, and Y. Sagiv. XSEarch: A Semantic Search Engine for XML. In *Proc. of 29th International Conference on Very Large Data Bases*, pages 45–56. Morgan Kaufmann, Sep. 2003.

[9] N. Gövert, N. Fuhr, M. Abolhassani, and K. Großjohann. Content-Oriented XML Retrieval with HyREX. In *Proc. of the First Workshop of the Initiative for the Evaluation of XML Retrieval*, pages 26–32. ERCIM, Mar. 2003.

[10] N. Gövert, G. Kazai, N. Fuhr, and M. Lalmas. Evaluating the Effectiveness of Content-oriented XML Retrieval. Technical report, University of Dortmund, Computer Science 6, Sep. 2003.

[11] L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram. XRANK: Ranked Keyword Search over XML Documents. In *Proc. of the 2003 ACM SIGMOD International Conference on Management of Data*, pages 16–27. ACM Press, June 2003.

[12] K. Hatano, H. Kinutani, M. Watanabe, M. Yoshikawa, and S. Uemura. Determining the Unit of Retrieval Results for XML Documents. In *Proc. of the First Workshop of the Initiative for the Evaluation of XML Retrieval*, pages 57–64. ERCIM, Mar. 2003.

[13] V. Hristidis, Y. Papakonstantinou, and A. Balmin. Keyword Proximity Search on XML Graphs. In *Proc. of the 19th International Conference on Data Engineering*, pages 367–378. IEEE CS Press, Mar. 2003.

[14] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. XML Retrieval: What to Retrieve? In *Proc. of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 409–410. ACM Press, Jul./Aug. 2003.

[15] M. Kaszkiel and J. Zobel. Passage Retrieval Revisited. In *Proc. of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185. ACM Press, July 1997.

[16] J. List and A. de Vries. CWI at INEX2002. In *Proc. of the First Workshop of the Initiative for the Evaluation of XML Retrieval*, pages 133–140. ERCIM, Mar. 2003.

[17] G. Navarro and R. Baeza-Yates. Proximal Nodes: A Model to Query Document Databases by Content and Structure. *ACM Transactions on Information Systems*, 15(4):400–435, Oct. 1997.

[18] A. Vries, D. Hiemstra, G. Kazai, J. Kamps, J. Thom, K. Hatano, N. Fuhr, N. Gövert, P. Ogilvie, T. Röllenke, and Y. Mass. Evaluation Metrics Workgroup's Report. `http://qmir.dcs.qmul.ac.uk/inex/Slides/wg_metrics.pdf`, Dec. 2002.