Topic Field Selection and Smoothing for XML Retrieval

Jaap Kamps   Maarten de Rijke   Börkur Sigurbjörnsson

Language & Inference Technology Group, ILLC, University of Amsterdam
Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands
{kamps, mdr, borkur}@science.uva.nl

ABSTRACT

Information retrieval from XML documents offers an opportunity to go below the document level in search of relevant information, making any element of an XML document a retrievable unit. We consider two dimensions along which we compare this element retrieval task with the traditional document retrieval task. We investigate how different topic representations and language model smoothing approaches affect the performance of the two tasks. We evaluate our ideas against the INEX 2002 XML retrieval test-suite.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Experimentation

Keywords

XML retrieval, language models, smoothing, topic representation

1. INTRODUCTION

XML documents differ from plain text documents. The latter, contain only plain text and they themselves are the natural unit of retrieval. XML documents, in contrast, are divided into a hierarchy of text objects, each of which could in principle be returned in response to a query. It is thus tempting to try to go below the document level and focus on retrieving document fragments that provide exhaustive yet concise answers to the users' information need.

In this paper we report on ongoing work aimed at comparing two XML retrieval tasks: XML document retrieval (return whole XML documents in response to an information need) and XML element retrieval (return focused elements only). Thus, our main question in this paper is the following:

Aim 1 How is XML element retrieval different from XML document retrieval?

In our comparison we focus on two aspects, closely related to the fact that XML elements and XML documents may vary widely in length: *topic field selection* and *language model smoothing*.

It is known that the use of additional topic fields from a test collection may affect retrieval effectiveness (see the related work section below). In particular, for adhoc retrieval the use of additional (longer) topic fields tends to increase performance, whereas for retrieval tasks that aim to retrieve sentences or other very small units, the use of longer topic representations tends to hurt performance. In principle, XML elements can range in length from very short (e.g., a single word) to the whole document. This, then, gives rise to the second of our main aims in this paper:

Aim 2 How does topic field selection affect the two tasks?

In recent years, language modeling approaches to information retrieval have attracted a lot of attention [20, 10, 16]. Language models are attractive because of their foundations in statistical theory, the great deal of complementary work on language modeling in speech recognition and natural language processing, and the fact that very simple language modeling retrieval methods have performed quite well empirically. The basic idea of these approaches is to estimate a language model for each document, and then rank documents by the likelihood of the query according to the estimated language model. Since document language models may suffer from inaccuracy due to data sparseness, a core issue in language modeling is *smoothing*. Smoothing refers to adjusting the maximum likelihood estimator for the document language model by, for example, combining it with a collection language model. The retrieval performance is generally sensitive to the smoothing parameters. Earlier studies in adhoc retrieval have found that for shorter queries the Jelinek-Mercer method works well with less smoothing (i.e., more weight is given to the document language model), while long queries require more smoothing (i.e., more weight is given to the collection language model) [23]. How do these findings carry over to the setting of XML document or element retrieval, and how are they influenced by the choice of topic fields? More generally, we have our third aim:

Aim 3 How does smoothing affect the two XML retrieval tasks?

To answer the questions raised above, we use the INEX test collection. The INitiative for the Evaluation of XML retrieval (INEX) was launched in 2002 to assess the effectiveness of retrieval methods for XML document and element retrieval [11]. The collection contains two kinds of topics. Content-only topics (CO) are traditional IR topics written in natural language. Content-and-structure

DIR 2003 4th Dutch-Belgian Information Retrieval Workshop, Amsterdam, The Netherlands

topics (CAS) are a mixture of natural language requirements and structural constraints. In our experiments we used the CO topics and their assessments; for a successful approach to the CAS topics, see [3]. INEX CO topics are divided into four fields, title, description, narrative and keywords. We used these fields, independently or in combinations, to create different topic representations.

The remainder of this paper is organized as follows. In Section 2 we discuss related work. Our experimental setup is described in Section 3 and Section 4 describes the experimental results. Conclusions and future work is discussed in Section 5.

2. RELATED WORK

Effective representation of users' information needs is an important issue, with a long history of experimental work. Many groups have performed comparative analysis in which a system's performance on one version of a topic is compared against its performance on another version; see, e.g., the annual TREC proceedings where this has proved to be a recurring theme. These types of investigation were greatly facilitated by the fact that, from the start, the TREC organizers decided to provide "user need" statements rather than more traditional queries [8]. While the topic fields varied somewhat from year to year during the early history of TREC, the current format has been in place for some time now; it has a short title field, a one-sentence description field, and a narrative field that is aimed at providing a complete description of document relevance for the assessors. This standard has been copied by many other tasks (such as multilingual retrieval, novelty) and by other evaluation exercises, including NTCIR [19], CLEF [4], and, to a large extent, INEX.

It is commonly taken for granted that longer statements of an information need generally result in improvements in retrieval effectiveness over shorter statements. This appears to be valid not just for adhoc retrieval, but for other adhoc-like tasks too [1], and not just for English but for many other languages as well (see e.g., [1, 17]). For some tasks, however, the statement does not seem to be valid. For instance, for the 2002 edition of the novelty task, one of the tasks the participants had to carry out, was to return relevant *sentences*, not documents, for a given information need. Several teams reported that using only the title field resulted in the best retrieval performance [9]. Shorter topic representations seem to be more effective for the novelty task because the "documents" are very short: long topic representations seem to cause lots of topic drift in this case.

What does this suggest for topic representation and XML retrieval? The XML element length distribution is very different from the XML document length distribution. Furthermore, in XML element retrieval there is a bias toward retrieval of large elements [12]. This prompts the question what type of topic representation is most suitable for XML retrieval. If we want to retrieve mostly (full-blown) articles, we should go for long topic representations. But if we mainly want to retrieve very short elements, the experience from the novelty task seems to suggest that short topic representations are to be preferred.

Choosing one topic representation over another may require a number of changes to a retrieval system's settings. Smoothing is one of the core issues in language modeling; it adjusts the maximum likelihood estimator so as to correct the inaccuracy due to data sparseness. The retrieval performance is generally sensitive to the smoothing parameters. The appropriate amount of smoothing has been found to be dependent on the topic representation [23]. The length of the topic representation has an impact on the optimal amount of smoothing. For all of these reasons it is interesting to see what the impact of smoothing is for the two XML retrieval tasks.

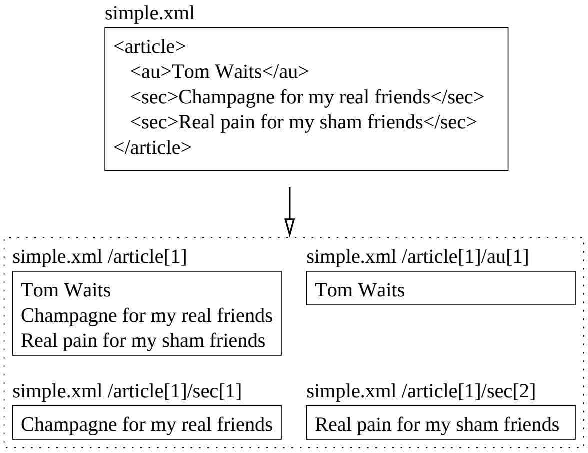
Smoothing is also task dependent. Language models for adhoc retrieval, and other tasks that are assessed in terms of mean average precision scores, tend to perform better if much smoothing is done [13, 10]. On the other hand, language models for high precision tasks such as web retrieval tasks seem to perform better if very little smoothing is applied [14]. With XML element retrieval we seem to be in a mixed situation: while it is assessed in terms of mean average precision, it can be thought of as a high precision retrieval task.

Much attention has been given to passage retrieval in the information retrieval community. The work has mainly focused on the use of passages to improve document retrieval [21, 2, 15]. Assessments have traditionally been performed on the document level, but not at the level of passages. Hence the evaluation of the passage retrieval is actually done at the document level. In [18] this approach has been adopted to XML retrieval; scores for individual XML elements are used to improve document retrieval in an SGML collection. These tasks are different from the XML element retrieval task discussed in this paper: the INEX collection provides assessments done directly on the element level. Hence the retrieval of XML elements proper is evaluated directly.

3. EXPERIMENTAL SETUP

We evaluate our ideas against the INEX 2002 XML information retrieval test-suite [7]. The INEX 2002 collection contains over 12,000 articles (consisting of nearly 7,000,000 elements) from 21 IEEE Computer Society journals, with layout marked up with XML tags. The collection contains around 170 different tag-names, representing units as diverse as complete articles `<article>`, sections `<sec>`, paragraphs `<p>` and italics font `<it>`.

To evaluate the two XML retrieval tasks, document and element retrieval, we need two types of indexes.



```
<article>
  <au>Tom Waits</au>
  <sec>Champagne for my real friends</sec>
  <sec>Real pain for my sham friends</sec>
</article>
```

simple.xml /article[1]
Tom Waits
Champagne for my real friends
Real pain for my sham friends
simple.xml /article[1]/au[1]
Tom Waits
simple.xml /article[1]/sec[1]
Champagne for my real friends
simple.xml /article[1]/sec[2]
Real pain for my sham friends

Figure 1: Simplified figure of how an XML document is split up into overlapping indexing units.

Element index Here, each element of an XML document is an indexing unit. For each element, all text nested within the element (including its descendants) is indexed (See Figure 1). This results in an overlapping element index, since the text nested at depth *n* is indexed as part of *n* different units.

Document index A fraction of the element index where only elements with a location path of depth 1 are considered (Such as the element with path `/article[1]` in Figure 1.

No stemming was applied to the indexes but we did lower-casing and stop-words were removed.

In our experiments we used the 23 CO topics that come with the INEX collection. INEX topics are divided into four fields: *title*, a short 2-3 word version of the topic statement; *description*, a one sentence definition of an information need; *narrative*, an explanation of the topic statement in more detail; and *keywords*, synonyms or terms that are broader/narrower than those listed in the title and description [6] (p.179). We used these fields, independently or in combinations, to create 5 different topic representations.

T The terms from *title*.

TD The terms from *title* and *description*.

TDN The terms from *title*, *description* and *narrative*.

TDK The terms from *title*, *description* and *keywords*.

TDNK The terms from *title*, *description*, *narrative* and *keywords*.

As with the collection, we did not stem the topics but lower-cased and removed stop-words.

At INEX 2002, relevance was assessed at the element level. Elements were assessed on a two dimensional graded relevance scale, one for topic relevance and another for element coverage [6] (p.184). From the official relevance assessments we derived two assessment sets, one for each of the tasks we want to evaluate.

Document retrieval task For evaluating the document retrieval we considered a document relevant if it contains an element judged highly relevant with exact coverage.

Element retrieval task For evaluating the element retrieval task we considered an element relevant if it was judged highly relevant with exact coverage.

We used version 1.8 of the INEX 2002 relevance assessments. Evaluation was done using the `trec_eval` program. Our evaluation method for element retrieval is similar to the strict evaluation used at INEX 2002 [6].

All our retrieval runs used a multinomial language model, with single length prior and Jelinek-Mercer smoothing [10]. Our scoring formula for an indexing unit d is thus

$$\begin{aligned} s(d) &= \log\left(\sum_t tf(t,d)\right) \\ &+ \sum_{i=1}^{n} \log\left(1 + \frac{\lambda \cdot tf(t_i,d) \cdot \left(\sum_t df(t)\right)}{(1-\lambda) \cdot df(t_i) \cdot \left(\sum_t tf(t,d)\right)}\right) \end{aligned}$$

where $tf(t,d)$ is the frequency of term t in document d and $df(t)$ is the count of document in which term t occurs. We experimented with a range of λs in the interval $[0.05, 0.95]$. In this paper we devote special attention to two values of the smoothing parameter used frequently in the literature. First, the default value of λ in adhoc retrieval: 0.15 [13]. Second, for high precision tasks such as web retrieval a high value of λ is normally used, such as 0.90 [14].

4. EXPERIMENTAL RESULTS

In this section we report on the results of our experiments for the two XML retrieval tasks. In Section 4.1 we look at the XML document retrieval task, in Section 4.2 we look at the XML element retrieval task and in Section 4.3 we compare the results for the two tasks. Since the combination of title and description fields is the most common topic representation in adhoc retrieval tasks, we use it as the baseline in our numeric comparisons. For determining whether a difference between retrieval runs is statistically significant, we use the bootstrapping method [5, 22]. We take 100,000 re-samples and look for improvement at significance levels 0.95 (*); 0.99 (**); and 0.999 (***).

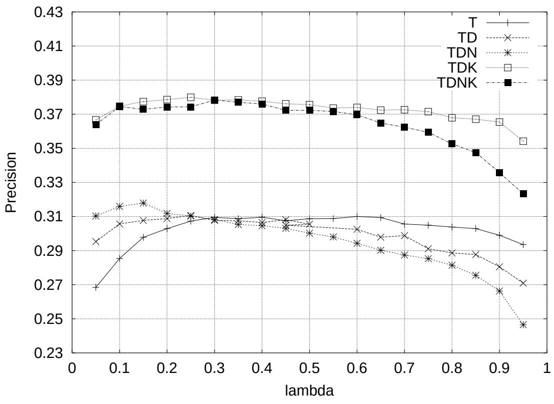


Figure 2: Mean average precision for document runs using different values for the smoothing parameter λ.

	$\lambda = 0.15$		$\lambda = 0.90$	
	MAP	% change	MAP	% change
TD (baseline)	0.3077	–	0.2805	–
T	0.2978	-3.2%	0.2990	+6.6%
TDN	0.3179	+3.3%	0.2663	-5.1%
TDK	**0.3774**	+22.7%***	**0.3654**	+30.3%***
TDNK	0.3729	+21.2%***	0.3358	+19.7%**

Table 1: MAP of document-runs using different query formats and different smoothing parameters.

4.1 XML document retrieval task

Figure 2 shows MAP scores of document runs for different values of the smoothing parameter λ. The first thing to notice is that the inclusion of keywords in the topic representation has the biggest positive impact on scoring (see TDN vs TDNK, and TD vs TDK). In Table 1 we compare the MAP scores of the topic representations for two values of the smoothing parameter λ. As the table shows, the runs using the keywords field are the only ones to improve significantly over the baseline.

We can also see that the topic representations containing the narrative field are the most sensitive to smoothing. This is not surprising since there may be various terms in the narrative that are not informative for the particular topic at hand. We see that the title-only-topics (T) is the only topic representation where less smoothing helps performance. Again, this is not surprising since the title-only queries do contain only good retrieval terms for the topic at hand. The T and TDK topic representations are the most stable over the range of values for the smoothing parameter. As before, the informativeness of all terms in the title and keywords fields is the most plausible explanation.

4.2 XML element retrieval task

Figure 3 shows MAP scores of element runs for different values of the smoothing parameter λ. For the element retrieval task, increased smoothing seems to hurt all topic representations, except for the title-only run. We again see that the queries including the keywords field give the best overall MAP score. Table 2 shows the

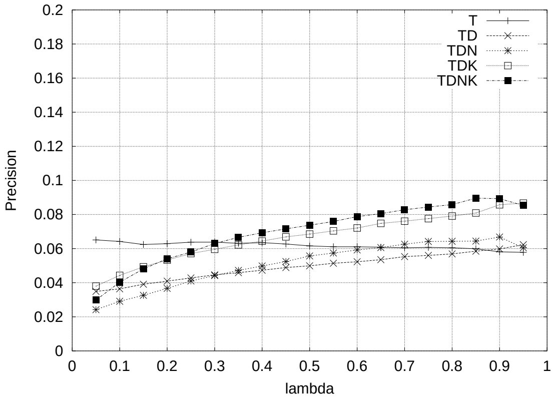


Figure 3: **Mean average precision for element runs using different values for the smoothing parameter λ.**

	$\lambda = 0.15$		$\lambda = 0.90$	
	MAP	% change	MAP	% change
TD (baseline)	0.0391	–	0.0598	–
T	**0.0624**	+59.6%	0.0581	-2.8%
TDN	0.0326	-16.6%	0.0668	+11.7%
TDK	0.0493	+26.1%	0.0857	+43.3%***
TDNK	0.0481	+23.0%	**0.0893**	+49.3%***

Table 2: **MAP of element-runs using different query formats and different smoothing parameters.**

MAP scores of element retrieval for two values of the smoothing parameter λ. It is clear from the table that the improvement of the keyword queries is only significant when little smoothing is done.

The amount of smoothing makes little impact on the short title-only topics. For all the longer queries, we see that higher values for the smoothing parameter do increase performance. The best smoothing settings for the XML element retrieval turn out to resemble those used for high precision tasks. This is somewhat surprising since the XML retrieval task is in modeled after a standard adhoc retrieval task where results are evaluated with MAP (i.e., average precision at all recall levels).

4.3 Documents vs. Elements

Looking at the results for the document and element retrieval tasks (Figures 2, 3 and 4), there is a striking difference between the performance of XML document retrieval and XML element retrieval. Document retrieval performs much better than element retrieval. This need not come as a surprise since we can look at the XML element retrieval task as a non-trivial extension of the XML document retrieval task. For the XML element retrieval task, given the set of relevant XML documents, we need to dive into each of the documents and retrieve the exact unit that made the document relevant.

We can also look at the ratio between the number of relevant documents and the number of documents in the collection and compare it to the ratio between the number of relevant elements in the collection and the total number of indexed elements in the collection (over all topics). (See Table 3)

$$\frac{rel.articles}{articles} = \frac{627}{12,107} \approx 0.0517$$

$$\frac{rel.elements}{elements} = \frac{1,394}{6,779,686} \approx 0.000206$$

	Count	Avg. len	Min len	Max. len
Document	12,107	3,234	24	21,333
Relevant	627	3,902	95	18,109
Element	6,779,686	29	1	21,333
Relevant	1,394	1.484	1	18,109

Table 3: **The count, average length, minimum length and maximum length of the set of documents, set of relevant documents, set of elements and set of relevant elements**

Finding the relevant elements seems to be a genuine needle-in-a-haystack problem.

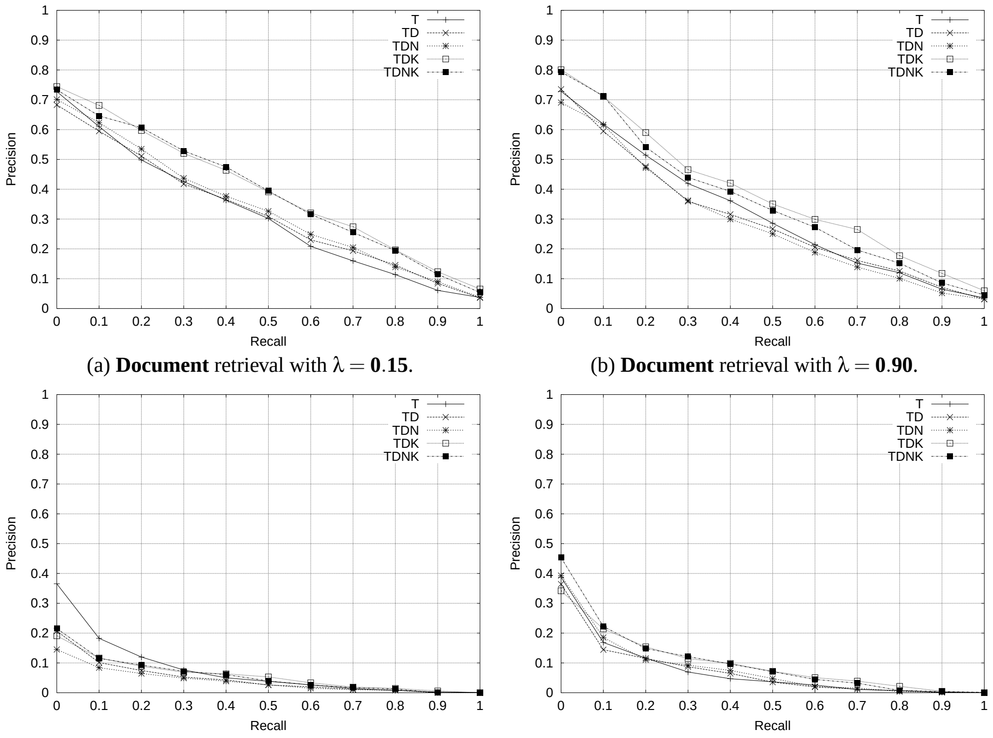
There are some similarities between the two tasks with respect to the impact of topic field selection. Adding terms from the keywords field leads to the biggest improvements. Longer topic representations will generally improve recall, but at the same time may hurt precision. Since the keyword field contains only terms that are informative for the topic at hand, we may expect little loss of precision. For terms in the other fields this need not be the case: both the description and narrative may contain terms that are not specific for the topic at hand. This is illustrated by the plots in Figure 4. The differences between the two tasks with respect to topic field selection largely depend on the used smoothing parameter.

The two tasks respond totally different to changes in the smoothing parameter λ. Much smoothing, i.e., a low value for λ, is the appropriate choice for the document retrieval task. This is in line with other adhoc retrieval experiments [23, 13, 10]. Little smoothing, i.e., a high value for λ, is the appropriate choice for the element retrieval task. We believe there are two factors working together toward providing the highest scoring for the element retrieval task. One is the high initial precision of the $\lambda = 0.9$ run; see Figure 4. Since it is extremely difficult to get high recall for this task, early precision is very important. Another factor is the size of the retrieved element. There is a serious variance in the length of elements in the collection and the average element length is low. However, assessors seem to have a strong bias toward larger elements [12]. Since we approach coordination level matching as $\lambda \rightarrow 1$ ([10] Appendix B), combining the long TDNK topic with little smoothing gives us a retrieval run that prefers elements which contain all the query terms, independent of whether they are informative or not. This may result in retrieval that has a similar bias toward larger elements as is present in the assessments.

Figure 5(c) shows the average length of retrieved elements for each of the values of the smoothing parameter. We can see a clear connection between the smoothing parameter and the average length of retrieved elements. For the longer topics (TD, TDN, TDK and TDNK), a higher value for λ causes larger elements to be retrieved on average. The opposite effect for the title only run is probably due to the fact that the length prior dominates in the scoring formula, as there are so few query terms. If we restrict our attention to the relevant elements retrieved we see the same tendency, but on a smaller scale (Figure 5(d)). Corresponding graphs for the document runs are shown in Figure 5(a) and (b). For comparison with the actual collection and assessment statistics see Table 3.

5. CONCLUSIONS

In Section 1, we introduced a number of research questions that motivated the experiments on which we reported in this paper. As for the second aim (how does topic field selection affect the two XML retrieval tasks?), we have seen that topic representations including keywords give the best MAP score for both tasks. Those



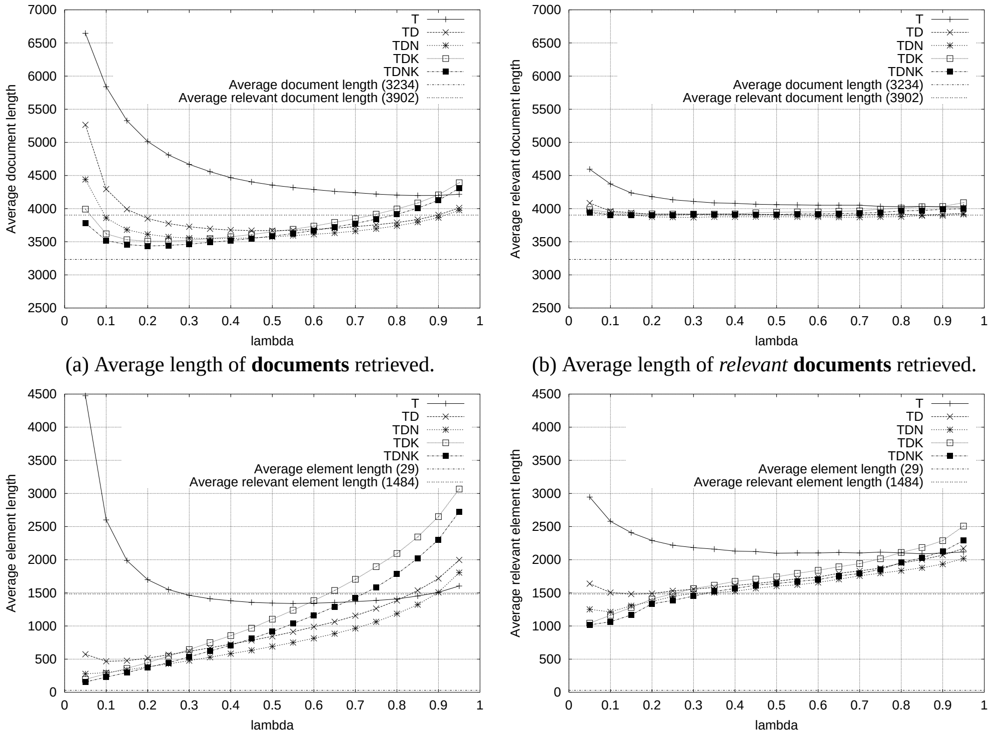
(a) **Document** retrieval with $\lambda = \mathbf{0.15}$.

(b) **Document** retrieval with $\lambda = \mathbf{0.90}$.

(c) **Element** retrieval with $\lambda = \mathbf{0.15}$.

(d) **Element** retrieval with $\lambda = \mathbf{0.90}$.

Figure 4: Precision-Recall curves for the different retrieval tasks, smoothing parameters and query formats.



(a) Average length of **documents** retrieved.

(b) Average length of *relevant* **documents** retrieved.

(c) Average length of **elements** retrieved.

(d) Average length of *relevant* **elements** retrieved.

Figure 5: Precision-Recall curves for the different retrieval tasks, smoothing parameters and query formats.

topic fields are the only one to give significantly better results than the TD topic baseline (for three of the four cases). Investigation of our third aim (how does smoothing affect the two XML retrieval tasks?) lead us to the finding that XML document retrieval reacts to smoothing in similar ways as other adhoc retrieval tasks. For XML element retrieval it turned out to be useful to use high-precision settings for the λ, even though the task is evaluated using a mean average prceision metric. For the XML document retrieval task, our best run uses the TDK topics and much smoothing (i.e., a low value of λ). For the XML element retrieval, our best run uses the TDNK topics and little smoothing (i.e., a high value of λ).

As an answer to the question expressed in our first aim (how is XML element retrieval different from XML document retrieval?), we have seen evidence that XML element retrieval is different from XML document retrieval. The tasks react radically differently to different amounts of smoothing. The difference with respect to changes in topic representation is more subtle, but we see that longer queries do always improve retrieval effectiveness (provided that the appropriate smoothing parameter is used).

It is not clear what effect the overlapping nature of the element index has on the statistics used to smooth the element language model. Indeed, it remains as future work to investigate different language models, such as an XML document language model, which can be used to smooth the element language model.

6. ACKNOWLEDGMENTS

Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 400-20-036 and 612.066.302. Maarten de Rijke was supported by grants from NWO, under project numbers 612-13-001, 365-20-005, 612.069.006, 612.-000.106, 220-80-001, 612.000.207, and 612.066.302.

7. REFERENCES

[1] J. Allan, J. Callan, B. Croft, L. Ballesteros, J. Broglio, J. Xu, and H. Shu. INQUERY at TREC-5. In *NIST Special Publication 500-238: The Fifth Text REtrieval Conference (TREC-5)*, pages 119–132, 1997.

[2] J. P. Callan. Passage-level evidence in document retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 302–310. Springer-Verlag New York, Inc., 1994.

[3] D. Carmel, Y. S. Maarek, M. Mandelbrod, Y. Mass, and A. Soffer. Searching XML documents via XML fragments. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 151–158. ACM Press, 2003.

[4] CLEF. Cross-Language Evaluation Forum, 2003. http://www.clef-campaign.org.

[5] B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

[6] N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas, editors. *Proceedings of the First Workshop of the Initiaitve for the Evaluation of XML Retrieval (INEX)*. ERCIM, 2003.

[7] N. Gövert and G. Kazai. Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002. In Fuhr et al. [6], pages 1–17.

[8] D. Harman. Overview of the First Text REtrieval Conference (TREC-1). In *NIST Special Publication 500-207: The First Text REtrieval Conference (TREC-1)*, pages 1–20, 1993.

[9] D. Harman. Overview of the TREC 2002 Novelty Track. In *NIST Special Publication: SP 500-251 The Eleventh Text Retrieval Conference (TREC 2002)*, 2003.

[10] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, 2001.

[11] INEX. Initiative for the evaluation of XML retrieval, 2003. http://www.is.informatik.uni-duisburg.de/projects/inex03/.

[12] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. XML Retrieval: What to Retrieve? In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 409–410. ACM Press, 2003.

[13] W. Kraaij, R. Pohlmann, and D. Hiemstra. Twenty-One at TREC-8: using language technology for information retrieval. In E. M. Voorhees and D. K. Harman, editors, *The Eighth Text REtrieval Conference (TREC-8)*, pages 285–300. National Institute for Standards and Technology. NIST Special Publication 500-246, 2000.

[14] W. Kraaij and T. Westerveld. TNO-UT at TREC-9: How different are web documents? In E. M. Voorhees and D. K. Harman, editors, *The Ninth Text REtrieval Conference (TREC-9)*, pages 665–672. National Institute for Standards and Technology. NIST Special Publication 500-249, 2001.

[15] X. Liu and W. B. Croft. Passage retrieval based on language models. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 375–382. ACM Press, 2002.

[16] D. R. H. Miller, T. Leek, and R. M. Schwartz. A hidden markov model information retrieval system. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 214–221. ACM Press, 1999.

[17] C. Monz and M. de Rijke. Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German and Italian. In *Proceedings CLEF 2001*, LNCS. Springer, 2002.

[18] S. H. Myaeng, D.-H. Jang, M.-S. Kim, and Z.-C. Zhoo. A flexible model for retrieval of SGML documents. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 138–145. ACM Press, 1998.

[19] NTCIR. NII-NACSIS Test Collection for IR Systems, 2003. http://research.nii.ac.jp/ntcir/index-en.html.

[20] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM Press, 1998.

[21] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–58. ACM Press, 1993.

[22] J. Wilbur. Non-parametric significance tests of retrieval performance comparisons. *Journal of Information Science*, 20:270–284, 1994.

[23] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342. ACM Press, 2001.