Using Topic Shifts in XML Retrieval at INEX 2006

Elham Ashoori and Mounia Lalmas

Queen Mary, University of London
London, E1 4NS, UK
{elham,mounia}@dcs.qmul.ac.uk

Abstract. This paper describes the retrieval approaches used by Queen Mary, University of London in the INEX 2006 ad hoc track. In our participation, we mainly investigate element-specific smoothing method within the language modelling framework. We adjust the amount of smoothing required for each XML element depending on its number of topic shifts to provide a focused access to XML elements in the Wikipedia collection. We also investigate whether using non-uniform priors is beneficial for the ad hoc tasks.

1 Introduction

In this paper we describe the Queen Mary, University of London's participation in the INEX 2006 ad hoc track.

Content-oriented XML retrieval systems aim at supporting more precise access to XML repositories by retrieving XML document components (the so-called XML elements) instead of whole documents in response to users' queries. Therefore, in principle, XML elements of any granularity (for example a paragraph or the section enclosing it) are potential answers to a query, as long as they are relevant. However, the child element (paragraph) may be more focused on the topic than its parent element (the section), which may contain additional irrelevant content. In this case, the child element is a better element to retrieve than its parent element, because not only it is relevant to the query, but it is also specific to the query.

To score XML elements according to how exhaustive and specific they are to a given query, various sources of evidence have been exploited. These include the content, the logical structure represented by the XML mark-up and the length of XML elements (e.g., [6,7,5]). In this work, we consider a different source of evidence, the number of topic shifts in an XML element. Our motivation stems from the definition of a relevant element at the appropriate level of granularity in INEX, which is expressed in terms of the "quantity" of topics discussed within each element. We therefore propose to use the number of topic shifts in an XML element, to express the "quantity" of topics discussed in an element as a means to capture specificity. Next, to compare this new feature to element length, we follow the spirit of [5], and incorporate this feature within a language modeling

N. Fuhr, M. Lalmas, and A. Trotman (Eds.): INEX 2006, LNCS 4518, pp. 261–270, 2007.
© Springer-Verlag Berlin Heidelberg 2007

framework, and examine its effects on the Wikipedia collection, compared to element length, to estimate prior probability of relevance in XML retrieval. Our previous work on IEEE collection showed that for retrieving highly specific and highly exhaustive elements, using topic shifts as prior is useful in the framework of language modeling [2].

For our experiments, we have implemented retrieval approaches for ranking XML elements based on a statistical language modelling approach [4]. Language modelling approaches have shown satisfactory results in content-oriented XML retrieval (e.g. [5,8]). This approach allows us to combine "non-content" features of elements (or documents) (e.g. length, topic shifts) with the scoring mechanism. We incorporate this new source of evidence, the number of topic shifts, as prior probability of relevance in the framework of language modeling. We also incorporate the number of topic shifts in the smoothing process within this framework as a means to capture specificity.

For the Thorough task, we experiment with two different ways of smoothing, element-specific smoothing and fixed smoothing approaches within the language modeling framework. We also compare topic shifts to element length, by incorporating each of them as prior probability of relevance and examining their effects on the retrieval effectiveness. For the Focused task, we apply a post-filtering algorithm to remove overlapping elements from our Thorough runs. We investigate whether using element-specific smoothing is beneficial for the Focused task. We also examine the usefulness of non-uniform priors for the Focused task. For the All In Context task, we took our runs for the Focused task, reordered the first 1,500 elements in the list such that results from the same article are clustered together. For the Best In Context task, we investigate whether retrieving the most focused element in a relevant article as the best entry point is a useful approach.

The paper is organised as follows. In section 2, we define topic shifts and how we calculate it. Section 3 and 4 describe the methodology and the experimental setting used in our investigation. The experiments and results are discussed in Section 6. Section 7 concludes the paper.

2 Topic Shifts

In this section, we describe how we measure the number of topic shifts of the elements forming an XML document. For this purpose, both the logical structure and a semantic decomposition of the XML document are needed. Whereas the logical structure of XML documents is readily available through their XML markup, their semantic decomposition needs to be extracted. To achieve that, we apply a topic segmentation algorithm based on lexical cohesion, TextTiling[1] [3], which has been successfully used in several IR applications. The underlying assumption of topic segmentation algorithms based on lexical cohesion, is that a change in vocabulary signifies that a topic shift occurs. This results in topic shifts being detected by examining the lexical similarity of adjacent text

[1] http://elib.cs.berkeley.edu/src/texttiles/

segments. TextTiling is a linear segmentation algorithm that considers the discourse unit to correspond to a paragraph and therefore subdivides the text into multi-paragraph segments.

The semantic decomposition of an XML document is used as a basis to calculate the number of topic shifts in each XML element forming that document. We consider that a topic shift occurs (i) when one segment ends and another segment starts, or (ii) when the starting (ending) point of an XML element coincides with the starting (ending) point of a semantic segment.

The number of topic shifts in an XML element e in document is defined as:

$$score(e) := actual\_topic\_shifts(e) + 1 \tag{1}$$

where $actual\_topic\_shifts(e)$ are the actual occurrences of topic shifts in element e of the document. We are adding 1 to avoid zero values. For simplicity, when we refer to the number of topic shifts, we shall be referring to $score(e)$.

With the above definition, the larger the number of topic shifts – i.e. the larger the $score(e)$ – the more topics are discussed in the element, i.e the content of element is less focused with respect to the overall topic discussed in the element.

3 Retrieval Framework

For our experiments, we have implemented retrieval approaches for ranking XML elements based on a statistical language modelling approach [4]. We rank elements based on the likelihood for a query $q = (t_1, t_2, ..., t_n)$ to be generated from an element e as:

$$P(e|q) \propto P(e) * P(t_1, .., t_n|e) \tag{2}$$

where

$$P(t_1, .., t_n|e) = \prod_{i=1}^{n} (\lambda_e P(t_i|e) + (1 - \lambda_e) P(t_i|C)) \tag{3}$$

and

- t_i is a query term in q,
- $P(e)$ is the prior probability of relevance for element e,
- $P(t_i|e) = \frac{tf(t_i,e)}{\sum_t tf(t,e)}$ is the probability of generating the query term t_i from element e,
- $tf(t, e)$ is the number of occurrences of term t in element e,
- $P(t_i|C) = \frac{ef(t_i)}{\sum_t ef(t)}$ is the probability of query term t_i in the collection,
- $ef(t)$ is the total number of XML elements in which term t occurs, and
- λ_e (weight on the element language model) is a weighting parameter between [0,1] which is used in smoothing the element model with the collection model.

We experiment with two ways of smoothing. First, we set λ_e to 0.1 for all elements, referred as fixed smoothing. This value is close to the traditional setting for document retrieval (λ=0.15), which has shown satisfactory results [4].

Secondly, to accommodate for the specificity dimension, we propose to set λ_e, the amount of smoothing, to be proportional to the number of topic shifts in the element, referred as element-specific smoothing. The idea of incorporating topic shifts in this manner originates from the fact that if the number of topic shifts in an element is low and an element is relevant, then it is likely to contain less non-relevant information compared to the case where a high number of topic shifts exists (For a complete argument see [1]). We define the element-specific smoothing parameter, λ_e, to be inversely proportional to the number of topic shifts in element e:

$$\lambda_e = \frac{\lambda}{score(e)} \tag{4}$$

where λ is a constant parameter between [0,1]. We set λ to 0.1 in the experiments where we use the element-smoothing approach.

We experiment with two different prior probabilities of relevance $P(e)$. First, following the work of Kamps etal in [5], we define the prior probability of relevance to be proportional to the length of an element. We refer to it as **length prior**:

$$P(e) = \frac{\sum_t tf(t,e)}{\sum_e \sum_t tf(t,e)} \tag{5}$$

Second, we define the prior probability to be proportional to the number of topic shifts in an element. We refer to it as **topic shifts prior**:

$$P(e) = \frac{score(e)}{\sum_e score(e)} \tag{6}$$

We also compare these two approaches with a baseline using a uniform prior. The uniform prior gives all elements an equal prior probability of being relevant.

4 Retrieval Setting

For calculating the number of topic shifts in each XML element, our first step is to decompose the Wikipedia XML documents into semantic segments through the application of TextTiling. We consider the discourse units in TextTiling to correspond to *paragraph* XML elements. We considered paragraph elements to be the lowest possible level of granularity of a retrieval unit. For the remainder of the paper, when we refer to the XML elements considered in our investigation, we will mean the subset consisting of paragraph elements and of elements containing at least one paragraph element as a descendant element.

Accordingly, the generated semantic segments can only correspond to paragraph elements and to their ancestors. As TextTiling requires a text-only version

of a document, each XML document has all its tags removed and is decomposed by applying the algorithm to sequences of paragraphs. We set the TextTiling parameters to $W = 10$ and $K = 6$. As a heuristic $W * K$ is equal to the average paragraph length (in terms of the number of terms) [3].

After the application of TextTiling in the above data sets, we compute the number of topic shifts in elements.

In this work, only the title field of the CO queries is used. No stemming is applied. Elements with size smaller than 20 has been removed when indexing the Wikipedia collection. When we refer to the size or the length of an element, we mean the number of terms after removing stopwords. For each of the retrieval approaches, the top 1,500 ranked elements are returned as answers for each of the CO topics.

5 Evaluation

For all tasks, we use the official metrics of INEX 2006. Since we only index and retrieve elements in the paragraph level or above, using the filtered assessment set will not change the relative order of our approaches considerably. Therefore we only reports the results using the full assessment set. In addition we report results for the Focused task using the strict quantization function. The strict quantization function is used to evaluate XML retrieval methods with respect to their capability of retrieving highly specific elements (s=1).

6 Experiments

6.1 Thorough Task

For the Thorough task we experiment with two different ways of smoothing, element-specific smoothing and fixed smoothing approaches ($\lambda = 0.1$ for all elements) in the framework of language modeling. We also consider both length and the number of topic-shifts as prior in addition to the uniform prior probability of relevance. Therefore, we consider six retrieval approaches in our experiments. Table 1 shows the details of our retrieval approaches where those runs submitted to INEX 2006 are marked with $*$.

Table 1. Thorough Retrieval Approaches

Approach	Prior	Smoothing
Lm_T	uniform	fixed
Lm_ToicShiftsPrior_T*	topic shifts	fixed
Lm_LengthPrior_T*	length	fixed
Lm_TermWeighted_T	uniform	element-specific
Lm_ToicShiftsPrior_TermWeighted_T*	topic shifts	element-specific
Lm_LengthPrior_TermWeighted_T	length	element-specific

Table 2. Thorough retrieval task: Evaluation based on Mean Average effort precision (MAep), using generalized quantization function

Approach	*MAep*
Lm_T	0.0179
Lm_ToicShiftsPrior_T*	**0.0185**
Lm_LengthPrior_T*	0.0181
Lm_TermWeighted_T	0.0144
Lm_ToicShiftsPrior_TermWeighted_T*	0.0163
Lm_LengthPrior_TermWeighted_T	0.0168

Table 2 presents, the evaluation results for Mean Average effort precision (*MAep*) for the six retrieval approaches. Focusing on either fixed smoothing approaches or approaches using element-specific smoothing, we observe that, using either the length prior or the topic shifts prior leads to slightly improvements of performance.

Focussing on the approaches employing non-uniform priors, we observe that they perform comparably, but none of them considerably improves the retrieval effectiveness compared to uniform prior, when evaluated with *MAep*.

When comparing the results for the approaches using fixed smoothing and element-specific smoothing, we see that fixed smoothing approaches are more effective in terms of *MAep* when evaluated under the generalized case.

6.2 Focused Task

The INEX 2006 Focused task asks systems to find the most focused elements that satisfy an information need, without returning "overlapping" elements. We experiment with the same approaches as we discussed for the Thorough task, and remove Overlap by applying a post-filtering on the retrieved ranked list. We select the highest scored element from each of the paths. In case of two overlapping elements with the same relevance score, the child element is selected. Therefore, we consider six retrieval approaches in our experiments. Table 3 shows the details of our retrieval approaches where those runs submitted to INEX 2006 are marked with $*$.

Table 3. Focused Retrieval Approaches

Approach	Prior	Smoothing
Lm_F	uniform	fixed
Lm_ToicShiftsPrior_F*	topic shifts	fixed
Lm_LengthPrior_F*	length	fixed
Lm_TermWeighted_F	uniform	element-specific
Lm_ToicShiftsPrior_TermWeighted_F*	topic shifts	element-specific
Lm_LengthPrior_TermWeighted_F	length	element-specific

Table 4. Focused retrieval task: *MAep* and normalised eXtended Cumulated Gain (*nxCG*) at different cut-off

Approach	nxCG@5	nxCG@10	nxCG@25	nxCG@50	MAep
General					
Lm_F	0.3477	**0.3075**	**0.2368**	**0.1818**	**0.0392**
Lm_TopicShiftsPrior_F*	0.3429	0.2953	0.2223	0.1649	0.0365
Lm_LengthPrior_F	0.3386	0.2751	0.2038	0.1559	0.035
Lm_TermWeighted_F	**0.3532**	0.2983	0.2282	0.1704	0.0369
Lm_TopicShiftsPrior_TermWeighted_F*	0.3455	0.2965	0.2351	0.1774	0.0382
Lm_LengthPrior_TermWeighted_F*	0.3525	0.2957	0.2276	0.1708	0.0381
Strict					
Lm_F	0.2937	0.2578	0.1929	0.1447	0.025
Lm_TopicShiftsPrior_F*	0.2721	0.2225	0.1627	0.1188	0.0193
Lm_LengthPrior_F	0.2342	0.182	0.131	0.101	0.016
Lm_TermWeighted_F	**0.3117**	**0.2614**	0.1965	0.1444	**0.0262**
Lm_TopicShiftsPrior_TermWeighted_F*	0.3009	0.256	**0.1993**	**0.1475**	0.0261
Lm_LengthPrior_TermWeighted_F*	0.3045	0.2487	0.1872	0.1377	0.0245

The evaluation results with respect to the *MAep* and *nxCG* at four different early cut-off points (5, 10, 25, 50) are shown in Table 4. For both evaluations, both strict and generalised quantization functions are used.

Under the generalized case, with all early precision measures apart from *nxCG*@5 using uniform prior and fixed smoothing approaches is the most effective approach.

Under the strict case, when comparing the results for the approaches using fixed smoothing and element-specific smoothing, we see that element-specific smoothing approaches, the bottom three runs, are more effective at early precision. This observation indicates the potential use of element-specific smoothing for retrieving highly specific elements at the early ranks.

The results also suggest that using non-uniform prior is not beneficial for the Focused task.

6.3 All in Context

For the All In Context task, we took our runs for the Focused task, reordered the first 1,500 elements in the list such that results from the same article are clustered together. We aim at examining the capability of our approaches in locating the relevant content within the relevant articles. Table 5 shows the details of our retrieval approaches where those runs submitted to INEX 2006 are marked with $*$.

The evaluation results with respect to Mean Average Generalized Precision (*MAgP*) and Generalized Precision (*gP*) at four different early cutoff point (5, 10, 25, 50) are shown in Table 6. Under all the official metrics of INEX 2006 for this task, using length prior and fixed smoothing provides the most effective approach in locating the relevant content within the relevant article. Using length

prior leads to considerable improvement over the uniform prior for gP@5. For the other measures used for this task, using both length and topic shifts prior leads to slight improvements of performance compared to the uniform prior.

Table 5. All In Context Retrieval Approaches

Approach	Prior	Smoothing
Lm_F_Clustered_R	uniform	fixed
Lm_TopicShiftsPrior_F_Clustered_R*	topic shifts	fixed
Lm_LengthPrior_F_Clustered_R	length	fixed
Lm_TermWeighted_F_Clustered_R	uniform	element-specific
Lm_TopicShiftsPrior_TermWeighted_F_Clustered_R*	topic shifts	element-specific
Lm_LengthPrior_TermWeighted_F_Clustered_R*	length	element-specific

Table 6. All In Context retrieval task Mean Average Generalized Precision ($MAgP$) and Generalized Precision at early ranks (gP) different cut-off

Approach	gP@5	gP@10	gP@25	gP@50	MAgP
Lm_F_Clustered_R	0.2572	0.2308	0.1760	0.1259	0.1147
Lm_TopicShiftsPrior_F_Clustered_R*	0.2593	0.2318	0.1759	0.1262	0.1179
Lm_LengthPrior_F_Clustered_R	**0.2833**	**0.2386**	**0.1766**	**0.1278**	**0.1232**
Lm_TermWeighted_F_Clustered_R	0.2489	0.2221	0.1637	0.1128	0.1028
Lm_TopicShiftsPrior_TermWeighted_F_Clustered_R*	0.2582	0.2270	0.1692	0.1193	0.1084
Lm_LengthPrior_TermWeighted_F_Clustered_R*	0.2595	0.2254	0.1702	0.1188	0.1105

6.4 Best in Context

The INEX 2006 Best In Context task asks systems to find the XML elements that corresponds to the best entry points to read articles. For the Best In Context task, we examine whether the most focused element in a relevant document is a good choice for the best entry point in a relevant article. For this task we took our official runs for the Focused task, and return for each article, the element with the maximum score as the best entry point. Table 7 shows the details of our official retrieval approaches. The last run marked with Δ is slightly different from Lm_TopicShiftsPrior_TermWeighted_F_B; such that in overlap-removal phase, in case of two overlapping elements with the same relevance score, the parent element is selected.

Table 7. Best in Context Retrieval Approaches

Approach	Prior	Smoothing
Lm_TopicShiftsPrior_F_B*	topic shifts	fixed
Lm_TopicShiftsPrior_TermWeighted_F_B*	topic shifts	element-specific
Lm_TopicShiftsPrior_TermWeighted_Foarent_B*Δ	topic shifts	element-specific

We report the INEX 2006 official results using the EPRUM-BEP-Exh-BEPDistance and BEPD metrics at five different values for A (0.01, 0.1, 1, 10, 100) as shown

in table 8. Low values of A (e.g. 0,1) favour runs that return elements very close to a best entry point.

Comparing the results of our runs, using fixed smoothing is the most effective runs for both metrics and for all values of A except at $A = 0.01$ where the approach based on element-specific smoothing outperformed. When evaluating these runs with EPRUM-BEP-Exh-BEPDistance at $A = 0.01$ (*low value*), our runs ranked very high among all participants. This shows that element-specific smoothing is useful at returning the elements very close to a best entry point in relevant articles.

Table 8. Best In Context task: EPRUM-BEP-Exh-BEPDistance and BEPD metrics

Approach	A=0.01	A=0.1	A=1	A=10	A=100
EPRUM-BEP-Exh-BEPDistance					
Lm_TopicShiftsPrior_F_B	0.0314	**0.0468**	**0.0735**	**0.1284**	**0.2024**
Lm_TopicShiftsPrior_TermWeighted_F_B	**0.0325**	0.0410	0.0610	0.1134	0.1855
Lm_TopicShiftsPrior_TermWeighted_Fparent_B	0.0300	0.0393	0.0607	0.1134	0.1855
BEPD					
Lm_TopicShiftsPrior_F_B	0.1129	**0.1611**	**0.2536**	**0.4092**	**0.5760**
Lm_TopicShiftsPrior_TermWeighted_F_B	**0.1259**	0.1591	0.2315	0.3815	0.5490
Lm_TopicShiftsPrior_TermWeighted_Fparent_B	0.1201	0.1587	0.2344	0.3836	0.5494

7 Discussion and Summary

This paper describes the retrieval approaches used by Queen Mary, University of London in the INEX 2006 ad hoc track. We participated in all four ad hoc track tasks. In this work, we experimented with two different ways of smoothing, fixed smoothing and element-specific smoothing approaches within the language modeling framework. We also investigated whether using non-uniform priors is beneficial for the ad hoc tasks in the Wikipedia collection. Our main findings are the following:

- Our results suggest that the the fixed smoothing approach is useful in several cases: (i) for the Thorough task, in terms of *MAep* and under the generalized quantization function, (ii) for the Focused task, for all early precision measures apart from $nxCG@5$ and under the generalized quantization function, (iii) for the All In Context task, in locating the relevant content within the relevant article with all the measures used for this task, (iv) for the Best In Context task, at finding the best entry point in the relevant elements for the values of $A > 0.01$ (we are looking for the elements very close to a best entry point when $A = 0.01$).
- For the element-specific smoothing, we used the number of topic shifts in the smoothing process. The idea of incorporating topic shifts in the element-specific smoothing approach originated from the fact that if the number of topic shifts in an element is low and an element is relevant, then it is likely to

contain less non-relevant information compared to the case with high number of topic shifts. Therefore, in this way of smoothing, in fact, we reward the presence of a query term in an element with a lower number of topic shifts (a more specific element). This means that we are capturing specificity with the number of topic shifts. Our results suggest that element-specific approach is useful in the following cases: (i) for the Focused task, in finding the highly specific elements at the early ranks, (ii) for the Best In Context, in finding the elements very close to a best entry point in relevant documents, i.e., for $A = 0.01$. These results indicate that the number of topic shifts is a useful evidence, as it seems to capture the specificity dimension of relevance.

– Finally, we observed that in general, using non-uniform prior is slightly beneficial for the Thorough and All In Context tasks, but not beneficial for the Focused task.

References

1. Ashoori, E., Lalmas, M.: Using topic shifts for focussed access to XML repositories. In: Advances in Information Retrieval: Proceedings of the 29th European Conference on IR Research (ECIR) (April 2007)
2. Ashoori, E., Lalmas, M., Tsikrika, T.: Examining Topic Shifts in Content-Oriented XML Retrieval, submitted (2006)
3. Hearst, M.A.: Multi-paragraph segmentation of expository text. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pp. 9–16 (1994)
4. Hiemstra, D.: Using Language Models for Information Retrieval. Phd thesis, University of Twente (2001)
5. Kamps, J., de Rijke, M., Sigurbjörnsson, B.: The importance of length normalization for XML retrieval. Information Retrieval 8(4), 631–654 (2005)
6. Mass, Y., Mandelbrod, M.: Using the inex environment as a test bed for various user models for XML retrieval. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 187–195. Springer, Heidelberg (2006)
7. Ogilvie, P., Callan, J.: Hierarchical language models for XML component retrieval. In: Fuhr, N., Lalmas, M., Malik, S., Szlávik, Z. (eds.) INEX 2004. LNCS, vol. 3493, pp. 224–237. Springer, Heidelberg (2005)
8. Ramirez, G., Westerveld, T., de Vries, A.P.: Using structural relationships for focused XML retrieval. In: Larsen, H.L., Pasi, G., Ortiz-Arroyo, D., Andreasen, T., Christiansen, H. (eds.) FQAS 2006. LNCS (LNAI), vol. 4027, pp. 147–158. Springer, Heidelberg (2006)