On the Relation between Relevant Passages and XML Document Structure

Jaap Kamps[1,2] Marijn Koolen[1]

[1] Archives and Information Studies, University of Amsterdam
[2] ISLA, Informatics Institute, University of Amsterdam

ABSTRACT

Whereas traditional document retrieval methods always return whole atomic documents as results, focused retrieval methods aim to provide more direct access to the relevant information by zooming in on those parts of the document that contain the relevant text. The main aim of this paper is to investigate how relevant text inside a document relates to the document structure. We analyze the INEX 2006 assessments, where topic assessors were asked to mark in yellow all and only relevant text, in relation to the underlying document structure of English Wikipedia pages transformed into XML.

Our main findings are: First, although relevant passages are typically small—with a median length of a few sentences and a mean length of a paragraph—they have varying lengths and may cover any fraction of an article. Second, the document structure corresponds reasonably well to the relevant passages. Although the shortest element containing the relevant passages is twice as long on average, half of the passages are closely fitting an XML element (the passage covers 95-100% of the element). Third, in particular the start of a relevant passage tends to coincide with the start of an XML element.

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

Measurement, Experimentation

Keywords

Evaluation, Relevance, Passage Retrieval, XML Retrieval

1. INTRODUCTION

In focused retrieval, the task is to go beyond the document level and zoom in on only those parts of the document that contain relevant text. Focused retrieval dates back, at least, to the early days of passage retrieval [6]. As Salton et al. [6, p.49] put it:

> Large collections of full-text documents are now commonly used in automated information retrieval. When the stored document texts are long, the retrieval of complete documents may not be in the users' best interest. In such circumstances, efficient and effective retrieval results may be obtained by using passage retrieval strategies designed to retrieve text excerpts of varying size in response to statements of user interest.

Early passage retrieval approaches have been using either the document structure (sentences, paragraphs, sections, etc.), or arbitrary text windows of fixed length [1]. In particular, the use of document structure derived from SGML mark-up was pioneered in [9]. The early experimental results primarily confirmed the effectiveness of passage-level evidence for boosting document retrieval. Over the years, research in this area has forked off several approaches like passage retrieval, question answering and XML element retrieval. In question answering, returning short and to-the-point results is a firm requirement [8]. In XML element retrieval, the goal is to retrieve those XML elements that are relevant (i.e., discuss the topic of request exhaustively) but contain no non-relevant information (i.e. they are specific for the topic of request) [2].

To evaluate focused retrieval methods, we also require relevance assessments below the document level. A simple binary decision whether the document is relevant no longer suffices. Assessors have to indicate which parts of the document are relevant, or in the case of question answering whether the given answer is correct, and evaluation measures have to reflect how well a *retrieved* document part fits a *relevant* document part. During the INEX 2006 campaign [5] such sub-document assessments have been collected. The document collection consists of the English Wikipedia pages transformed into XML [4]. Topic assessors are asked to mark in yellow all and only relevant text in a pooled set of documents. The judges only view the rendered text, unaware of the precise underlying XML structure. As a result, the highlighted passages are elicited unobstructed by the XML document structure.

The main aim of this paper is to investigate how relevant text inside a document relates to the document structure. Recall from the above, passages have traditionally been defined using either the document structure (like the XML structure at INEX), or based on various windows of text (like the assessors' highlights). This prompts a number of questions:

- What is the length of relevant passages? What fraction

SIGIR 2007 Workshop on Focused Retrieval
July 27, 2007, Amsterdam, The Netherlands

Table 1: Length of relevant passages in the INEX 2006 adhoc assessments.

	Min	Max	Median	Mean	Stdev
passage length	1	78,943	297	1,090	3,263
article length	96	234,461	4,528	9,485	12,962
article highlights	7	78,943	510	1,753	4,242
article ratio	0.0001	1.0000	0.1339	0.3160	0.3574

of the article is considered relevant?

- How well do the highlighted passages correspond to XML elements of the document structure?
- Since highlighted passages may span a range of elements, how do the passage boundaries correspond to XML element boundaries?

The adhoc task at INEX is to retrieve XML elements containing relevant text at the right level of granularity. The adequacy of the document structure to determine the unit of retrieval has been challenged in [7]. To study the value of the XML document structure to define retrieval results, INEX is allowing also arbitrary passage results in 2007. The analysis of this paper differs from the INEX retrieval tasks: rather than evaluating retrieval results in terms of their relevant or highlighted text, we investigate the highlighted passages as a whole directly.

2. ANALYSIS

We analyze the INEX 2006 adhoc retrieval assessments (v5-filtered) containing judgments for 114 topics (numbered 289-298, 300-366, 368-369, 371-376, 378-388, 390-392, 394-395, 399-407, 409-411, and 413). The assessors have assessed relevance by highlighting relevant text at the granularity of sentences. The assessment interface automatically merges consecutive highlighted passages. A passage's start and end point is identified by either XML element boundaries or character-offsets on the respective text nodes. First, we will look at the length of passages, both in absolute and relative terms. Second, we will investigate how highlighted passages relate to XML elements. Third, we will zoom in on the passages start and end points, and relate them to XML element boundaries.

2.1 Relevant Passage Length

We start by looking at the length of highlighted passages, both absolute and relative length, and want to find out characteristics of the relevant information inside articles. Table 1 shows the length of highlighted passages for the INEX 2006 adhoc topics. Over 114 topics, there are 9,086 passages in 5,648 articles (we restrict our analysis to these articles). Passages contain 1,090 characters on average (median 297), while relevant articles contain almost 10,000 characters on average (median 4,528). Since articles can have multiple relevant passage, the average length of relevant text per article is 1,753 characters, showing that these relevant articles have 1.6 relevant passages on average. Looking at the relative length of the highlights, we see that on average 31.60% of the relevant articles' text is highlighted (median 13.39%). The highlighted passages have a median length of a couple of sentences, and an average length of a paragraph.

We now look at the impact of the topic at hand on the length of the highlighted passage. Figure 1 shows the distribution of passage length over topics. Although most of the passages are very short, some topics contain quite a few passages that are over 10,000 characters in length. There is certainly no "fixed" passage length per topic. Moreover, there is variation in length of highlighted passages over topics, although also plotting the relevant article's length over topics (not shown) results in similar pattern.

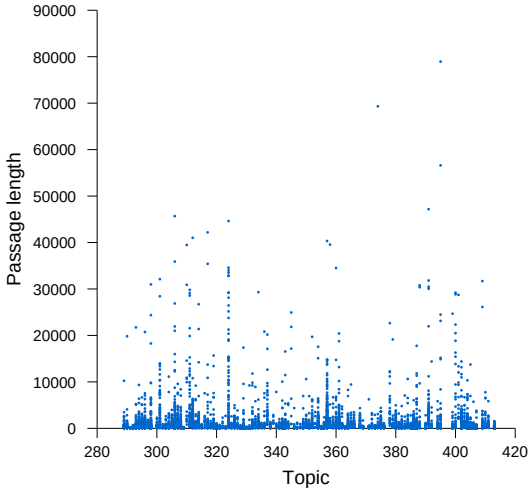

Figure 1: Length of highlighted passages over topics.

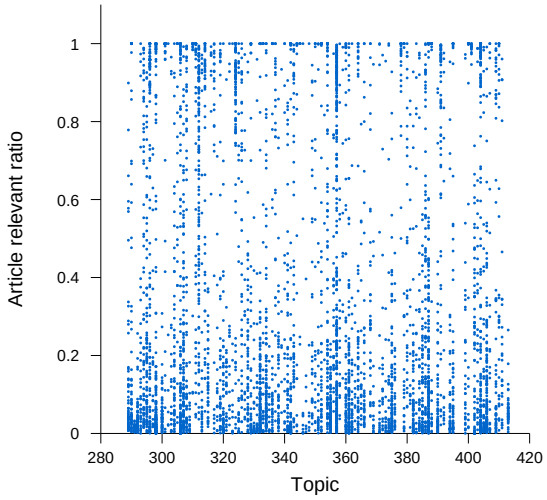

Figure 2: Fraction of the article that is highlighted over topics.

Since articles have substantial variation in length, we look at the relative length of the highlighted text. Figure 2 shows the fraction of articles that is highlighted over topics. What is most striking is the spread over the whole range. For many of the articles across most topics, only a small fraction (less than 20%) of the text is highlighted. Also, for many topics, there are a few articles that are wholly relevant. The density of the plot seems somewhat greater on the extremes.

Does the fraction of highlighted text depend on the length of the article? Figure 3 shows the fraction of articles that is highlighted over the length of the articles. Many of the Wikipedia articles are rather short, including many of the relevant articles. Most of the relevant articles are much shorter than 50,000 characters, and for most of the articles the relevance ratio is below 0.2, corresponding to Fig-

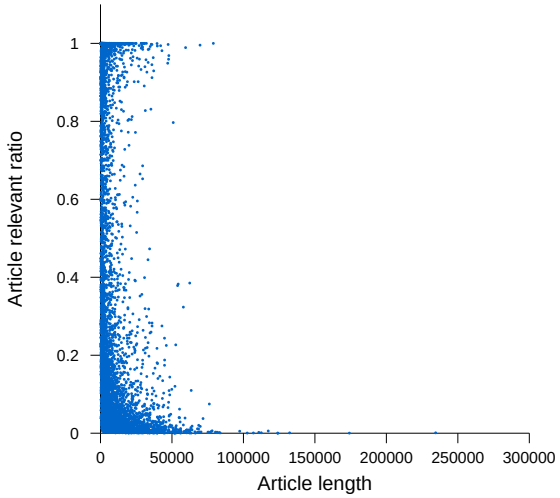

Figure 3: Article length versus highlighted fraction.

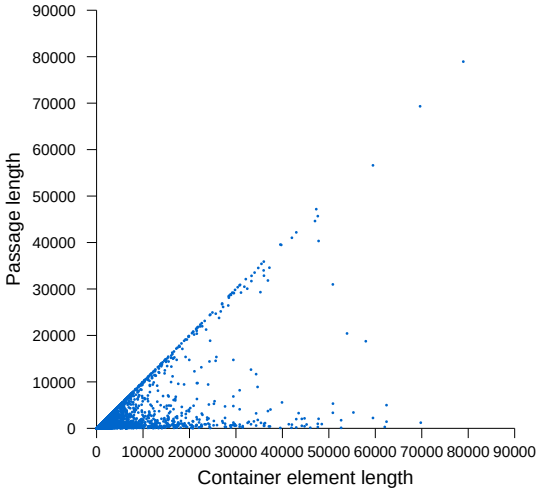

Figure 4: Passage length versus component element length.

Table 2: Length of passages and container elements.

	Min	Max	Median	Mean	Stdev
passage length	1	78,943	297	1,090	3,263
container length	1	78,943	620	2,348	5,525
container ratio	0.0009	1.0000	0.9730	0.7028	0.3637

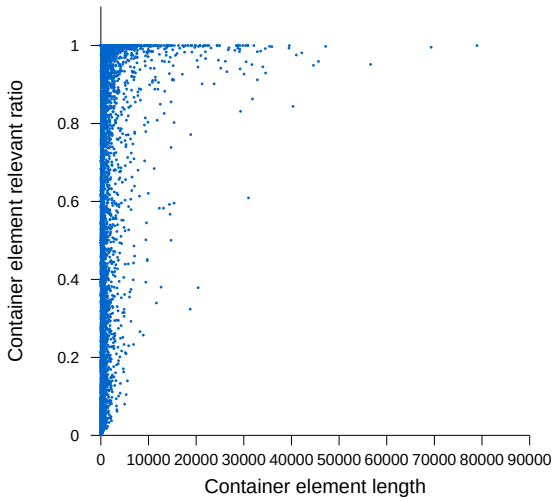

Figure 5: Fraction of container element that is highlighted.

ure 2. Above a relevance ratio of 0.2, the articles are spread more or less evenly over the relevance ratio scale, indicating that the relevant portion of an article varies greatly. This is rather surprising, as we would expect that longer articles have a smaller percentage of relevant text. Recall from the introduction that sub-document retrieval is motivated by the assumption that long documents only contain a relatively small fraction of relevant text.

Summarizing, our analysis showed that i) relevant passages are relatively short with a median length of a couple of sentences, and an average length of a paragraph; ii) there is no "fixed" length of relevant passages; iii) the highlighted text may cover any fraction of the article; and iv) the fraction of the article that is highlighted does not depend on the length of the article.

2.2 Relating Passages to Elements

We now relate the relevant passages to the document structure, and want to find out how well the highlighted passages correspond to XML elements of the document structure. From the article level, we now zoom in on the XML elements that contain relevant text. We use the notion of *container elements* to identify those elements that contain the whole relevant passage. More specifically, we will focus on the *shortest container elements*, i.e. the shortest element to contain the *whole* passage.

How long are the XML elements containing the passages? Table 2 gives some statistics on the length of passages and their container elements. We include the passage lengths again for comparison. The container elements have a mean length of 2,348 characters, and a median length of 620 characters. That is, the average container element is twice the length of the average passage. The minimum and maximum lengths are equal, meaning that both the shortest passage and the longest passage exactly fit their container element, i.e. the container contains only relevant text. This suggests that if we approximate the relevant passage by an XML element from the document structure, we retrieve in total twice the length of relevant text. The ratio of the container elements that is covered by the relevant passage, also shown in Table 2, is on average 70% but the median ratio is 97%. This suggests a reasonable fit between passages and their container elements.

In the previous section we saw that relevant passages vary widely in length. How does the length of the passages relate to the length of the container element? Figure 4 plots the passage length against the container element length. The diagonal axis shows the passages that exactly fit their container elements, and especially for longer passages the container element fits like a glove. The part below this diagonal axis is empty, as passages can never be longer than their container elements. The bulk of the passages is shorter than 10,000 characters, and here their containers are often substantially longer than the relevant passages. Looking at the same data from another angle, Figure 5 plots the ratio of container elements that is highlighted. This shows the same pattern: the longer containers tend to have higher relevance

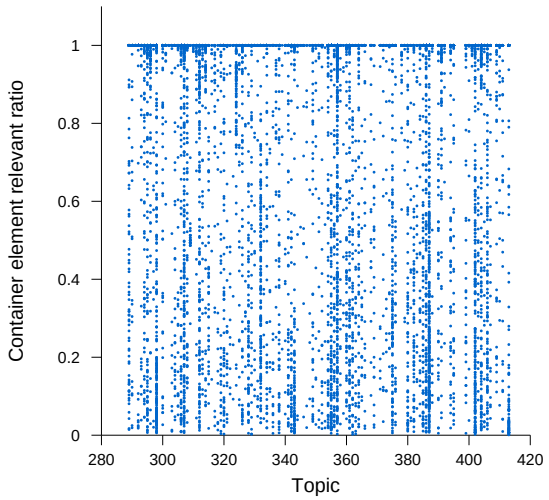

Figure 6: Fraction of container element that is highlighted over topics.

Table 3: Distribution of container elements over relevance ratio.

Ratio	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Frequency	419	755	656	467	432	375	315	247	288	424	4,705

Table 4: Container tag frequency and mean relevance ratio.

Tag	Frequency	Mean length	Mean ratio
⟨p⟩	2,761	558.7	0.7045
⟨body⟩	1,693	6,184.8	0.4213
⟨section⟩	1,424	2,453.6	0.6746
⟨item⟩	944	138.2	0.9248
⟨article⟩	724	7,009.6	0.8526
⟨normallist⟩	304	1,004.8	0.4667
⟨name⟩	270	21.4	1.0000
⟨collectionlink⟩	209	19.4	1.0000
⟨row⟩	180	62.0	0.7122
⟨caption⟩	174	93.7	0.9849

ratios. This is in itself no big surprise, since a long relevant passage spanning a range of elements is required for these long container elements.

Some of the topics provide hints of the type of XML element that is likely to be relevant. Does the topic at hand impact the relative fit of the container element? Figure 6 shows the relevance ratio of the container elements split over topics. For many topics, the number of container elements with smaller ratios is small, but there is great variation in relevance ratios over containers. The dark line at the top indicates that quite a number of relevant passage boundaries coincide with the container element boundaries. From the plots it is still not clear whether the number of containers with a relevance ratio of 1 is higher than the number of containers at lower relevance ratios. Table 3 shows the distribution of container elements over the different relevance ratios. In total, 4,705 relevant passages closely fit their container element, that is, half of the relevant passages (51.8%) cover 95–100% of the text of their container elements.

Finally, we investigate the correspondence between specific container element types and highlighted passages. Table 4 shows the tag names of the container elements, their frequencies, mean length, and the mean of their relevance ratios. The ⟨p⟩ element is the most frequent container of relevant passages and on average, 70% of these containers is relevant text. The ⟨body⟩ element is also very frequent but has a much lower relevance ratio (42%). The ⟨article⟩ element, somewhat surprisingly, has a much higher relevance ratio (85%), while it is only slightly longer than the ⟨body⟩ element. The ⟨article⟩ contains the ⟨body⟩ element and the elements ⟨name⟩ (the name of the Wikipedia article) and ⟨conversionwarning⟩. A plausible explanation is that if a large part of the article is relevant, the ⟨name⟩ of the page will be included in the passage highlighted by the assessor, resulting in ⟨article⟩ being the container element. If the ⟨name⟩ element is not highlighted, but different sections somewhere down the article are highlighted, the container element will be the ⟨body⟩. Other document structures that correspond well to highlighted passages are ⟨section⟩, ⟨item⟩, ⟨name⟩ and ⟨collectionlink⟩ elements.

Summarizing, our analysis above revealed mixed results for the correspondence between relevant passages and container elements (i.e., the shortest XML element containing the whole passage). On the one hand, the average container element is twice as long as the average passage. On the other hand, half of the passages have a closely fitting container element (the passage covers 95-100% of the element).

Table 5: Offsets of relevant passages.

	Min	Max	Median	Mean	Stdev
start element	0	10,723	0	62.74	317.68
end element	0	61,743	2	365.80	2,423.29
start container	0	47,510	1	252.90	1,344.91
end container	0	68,566	24	1,023.48	3,928.68

2.3 Passage and Element Boundaries

We now zoom even further in, and look at the relation between passage boundaries and element boundaries. We define two more notions, *start element* and *end element* as:

- *start element*: the XML element that directly contains the first highlighted character of the passage.
- *end element*: the XML element that directly contains the last highlighted character of the passage.

If the highlighted passage crosses no element boundaries (e.g., a passage from a single paragraph), the start and element elements coincide and are also the container element.

We look at where the highlighted passages start and end (character offset) in the document structure and within their container elements. Table 5 shows the offsets of highlighted passages for the INEX 2006 adhoc topics. First, we look at the closest XML element boundaries and see that the median offset in the start element is 0. Thus, at least half of the highlighted passages start at an XML element boundary. The much higher mean offset shows that the distribution is skewed. Nonetheless, the bulk of the passages start very close to the start element boundary. Second, the offset to the end of the end element is 2, showing that most the passages end at the boundary of the end element. The average is much higher, showing again a skewed distribution. Third, we look at the shortest XML element containing the whole passage and see that the median offset in the container element is 1, indicating that many of the container elements

are also start elements. Fourth, the median offset to the end of the container elements is 24, showing that most of the passages end some distance before the end the container element.

Summarizing, the correspondence between the relevant passages and document structure is particularly strong at the passages' start points: relevant passages start at an element boundary.

3. CONCLUSIONS

In focused retrieval the aim is to retrieve only those parts of a document that contain relevant text and no non-relevant text. In XML retrieval the XML structure of documents is exploited to locate relevant elements and use their boundaries as passage boundaries. In this paper we have investigated how well these XML element boundaries correspond to the boundaries of relevant passages in the INEX 2006 adhoc assessments.

Our first question was:

- What is the length of relevant passages? What fraction of the article is considered relevant?

The data show that most relevant passages are rather short, less than 1,000 characters, but there is a great variety over topics, and there seems to be no 'fixed' passage length and there is no relation between passage length and article length, and therefore no clear answer on what fraction of an article is considered relevant.

The second question was:

- How well do the highlighted passages correspond to the XML elements of the document structure?

The average length of the shortest element containing the highlighted passage is twice as long as the average passage length, but half of these container elements are a close fit to the passage (95-100% of their content being relevant text). Document structures that correspond naturally to highlighted passages are paragraphs, sections, list-items, titles and the whole article itself. However, even though these structures correspond reasonably well to highlighted passages, there is large variation over passages, articles and topics.

Our last questions was:

- Since highlighted passages may span a range of elements, how do the passage boundaries correspond to XML element boundaries?

The start of the passage often corresponds with the first character of the "start" element and the container element. The end of the passage corresponds well to the last character of the "end" element, and is at some distance from the end of the container element.

There are, as always, various limitations to the analysis provided. First, there is an obvious impact of the particular document structure of the collection. Wikipedia is an encyclopedia, with a highly organized structure, and created by a multitude of writers and editors. The generated XML encoding is based on the simple Wiki-syntax, and of course depends the particular writing style—how well is the particular article textually structured? and how well does this correspond to the sectioning structure? Second there is an obvious impact of relevance assessor and the assessment interface. Does a judge highlight the best text in the article's context, or judge relevance on equal grounds throughout the whole collection?

What do we learn from the analysis in terms of the retrieval approaches? First, the short length of the typical relevant passage seems to suggest retrieving fixed window passages, but the variation in length of passages and coverage of the article seems to suggest a flexible unit of retrieval such as XML elements. Second, the fact that half of the passages fit closely with an XML element seems to support retrieving XML elements, but the fact that the corresponding elements are twice the length of the relevant passage seems to support passages results. Third, the start of a relevant passage tends to coincide with the start of an XML element, so if we assume results are displayed in the context of the article, retrieval of XML elements seems a good approach. Although also fixed window passage retrieval proved an effective approach to find hot-spots inside articles [3]. In short, there is mixed support for both retrieving elements of the document structure and for retrieving arbitrary passages. We look forward to the retrieval experiments at INEX 2007 to help determine what approaches turn out to be more effective in practice.

Acknowledgments

Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO, grants # 612.066.513, 639.072.601, and 640.001.501), and by the E.U.'s 6th FP for RTD (project MultiMATCH contract IST-033104). Marijn Koolen was supported by NWO (# 640.001.501).

REFERENCES

[1] 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 NY, 1994.

[2] C. Clarke, J. Kamps, and M. Lalmas. INEX 2006 retrieval task and result submission specification. In N. Fuhr, M. Lalmas, and A. Trotman, editors, *INEX 2006 Workshop Pre-Proceedings*, pages 381–388, 2006.

[3] C. L. A. Clarke and E. L. Terra. Passage retrieval vs. document retrieval for factoid question answering. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 427–428. ACM Press, New York NY, 2003.

[4] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 40(1):64–69, June 2006.

[5] INEX. INitiative for the Evaluation of XML Retrieval, 2006. `http://inex.is.informatik.uni-duisburg.de/2006/`.

[6] 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, New York NY, 1993.

[7] A. Trotman and S. Geva. Passage retrieval and other XML-retrieval tasks. In A. Trotman and S. Geva, editors, *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology*, pages 43–50, 2006.

[8] E. M. Voorhees. Overview of the TREC 2001 question answering track. In *The Tenth Text REtrieval Conference (TREC 2001)*, pages 42–51. National Institute for Standards and Technology. NIST Special Publication 500-250, 2002.

[9] R. Wilkinson. Effective retrieval of structured documents. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 311–317. Springer-Verlag, New York NY, 1994.