Where to Start Reading a Textual XML Document?

Jaap Kamps[1,2] Marijn Koolen[1] Mounia Lalmas[3]

[1] Archives and Information Studies, University of Amsterdam
[2] ISLA, Informatics Institute, University of Amsterdam
[3] Department of Computer Science, Queen Mary, University of London

ABSTRACT

In structured information retrieval, the aim is to exploit document structure to retrieve relevant components, allowing the user to go straight to the relevant material. This paper looks at the so-called best entry points (BEPs), which are intended to give the user the best starting point to access the relevant information in the document. We examine the relationship between BEPs and relevant components in the INEX 2006 ad hoc assessments. Our main findings are the following: First, although documents are short, assessors often choose the best entry point some distance from the start of the document. Second, many of the best entry points coincide with the first relevant character in relevant documents, showing a strong relation between the BEP and relevant text. Third, we find *browsing BEPs* in articles with a single relevant passages, and *container BEPs* or *context BEPs* in articles with more relevant passages.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries

General Terms: Measurement, Performance, Experimentation

Keywords: XML Retrieval, Element retrieval, Best entry point

1. INTRODUCTION

Focused structured document retrieval employs the concept of best entry point (BEP), which is intended to provide the best starting point to access the relevant information in the document [2]. In this paper, we examine the relationship between BEPs and relevant components in the INEX 2006 ad hoc assessments [3]. Earlier research on a collection of Shakespeare's plays used multiple BEPs due to the length of the plays [4]. Given that our collection consists of short and topically focused articles (derived from Wikipedia), we asked the assessors to provide only a single BEP for each article with relevant information.

INEX 2006 used an XML'ified collection of English Wikipedia pages, containing over 650,000 articles, 1,241 unique tags, and an average element depth of 4.8 [1]. Topic assessors were asked to find relevant passages by marking all and only the relevant text in yellow, and to point out the best place to start reading relevant information (BEP). The data consists of 111 topics, 5,308 relevant Wiki-pages, 5,483 BEPs, and 8,737 relevant passages. As Wikipedia pages are relatively short with relevant articles containing on average 9,000 characters, assessors might simply judge the start of the article to be the BEP. Hence, we want to know three things: What elements does the BEP correspond to, and what is the depth of these elements in the article? Where is the BEP with respect to the start of the article? Where is the BEP with respect to the relevant text?

Table 1: BEP tags and mean/median depth

Tag-name	Frequency	Mean	Median
⟨p⟩	1,652	3.83	4
⟨name⟩	983	2.00	2
⟨emph3⟩	613	3.34	3
⟨collectionlink⟩	576	5.44	5
⟨title⟩	439	4.56	4
⟨body⟩	352	2.00	2
⟨item⟩	189	5.22	5
⟨section⟩	122	3.53	3
⟨unknownlink⟩	86	5.85	3
⟨caption⟩	71	4.44	4

Table 2: Assessments statistics for the INEX 2006 ad hoc topics

	N	Min	Max	Median	Mean	Stdev
article length	5,483	99	234,460	4,405	9,343	12,937
BEP	5,483	0	113,320	556	3,090	6,856
FRC	5,481	1	113,320	477	2,938	6,659

2. BEST ENTRY POINTS IN INEX 2006

We processed the INEX 2006 assessment files and relevant Wikipedia articles, and computed the distances between the start of the article and the *best entry point* (BEP) and the *first relevant character* (FRC). All the distances are in character length.

Best Entry Point First, we look at which tags the BEP is placed. Table 1 shows the 10 most frequent tags, and their mean and median depth in the document structure. The most frequent BEP tag is paragraph (⟨p⟩), which is nested deeply in the document (at median depth 4). The ⟨p⟩ is also one of the most frequent tags in the collection. The second most frequent tag is ⟨name⟩, the main title of the Wiki-page at the start of the article. Since the structure was not shown explicitly to the assessor, the ⟨name⟩ tag also indicates that the assessor regards the start of the article as BEP, even though the whole ⟨article⟩ tag occurs relatively infrequently (42 times). The same holds for the ⟨title⟩ which is also the first content of ⟨section⟩. The depth of the elements varies between 2 (⟨name⟩) and 5 (⟨collectionlink⟩).

Second, we look specifically at how far into the article the BEP is placed. Table 2 (second row) shows the distance in characters between the start of the article and the BEP. What we see is that the BEP is a fair distance into the article (median distance 556, mean distance 3,090). The difference between median and mean distance signals that the distribution is skewed toward the start of the article. Comparing the BEP distance and the length of the article, we find a significant correlation of 0.66.

Third, we zoom even further in and look at where the BEP is placed relative to the whole article's length. Table 3 (second row) shows whether the BEP is in the first percentages of the article's

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

Table 3: Distribution of BEP and FRC at % of article length

% of article	1	2	3	4	5	6	7	8	9	10+
% BEPs	25.85	5.31	3.10	2.50	1.72	1.13	1.00	1.06	1.08	57.25
% FRCs	26.33	6.08	3.58	2.97	2.12	1.09	1.26	1.09	1.04	54.44

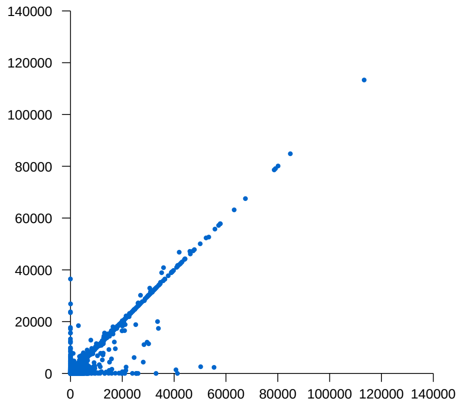

Figure 1: Distance of BEP (x-axis) to the FRC (y-axis)

length. We see that a quarter of the BEPs is placed in the first percent of the article, and over 40% in the first 10 percent of the article. BEPs do not necessarily appear at the start of the article, but are spread out through the whole article.

Summarizing, what we observe is that the majority of BEPs are placed inside the article. This clearly signals a preference for more focused starting points than the whole article, even in the case of the relatively short and single-faceted Wikipedia articles.

First Relevant Character We now look at the BEPs relative to the FRC. Table 2 (third row) shows where the first relevant character is located with respect to the start of the article. The pattern for the FRC is very similar to the BEP: The FRC also starts at a fair distance into the article (median distance 447, mean distance 2,938). Also the relative placement of the FRC in the article (third row in Table 3) shows a distribution very similar to the BEP.

We look at the relation between BEP and FRC directly. Figure 1 shows the location of the BEP set off against the location of the FRC. There is a very clear diagonal indicating that the BEP and FRC often coincide. Indeed, there is a highly significant correlation of 0.94 between the BEP and the FRC. The majority of BEPs (54.39%) coincide with the FRC. This is called *browsing BEPs*, i.e., the best entry point is the start of the first relevant object [2].

Summarizing, there is a clear relation between the BEP and the FRC: in the majority of cases the best place to start reading is exactly the beginning of relevant text in the article.

Number of Highlighted Passages We now break down the BEPs over the number of highlighted passages in the article. A relevant article has on average 1.59 relevant passages, although the majority of relevant articles has only a single highlighted passage. Table 4 shows where the BEP is placed before, on, or after the FRC. We see that the BEP coincides with the FRC in 67% of the articles that have only one relevant passage. However, in articles with multiple relevant passages, the BEP is most frequently placed before the FRC. The majority of BEPs for articles with multiple relevant passages are either *context BEPs*, i.e., appearing at the same structural level as the relevant objects, or *container BEPs*, i.e., higher level objects which contain several relevant objects [2]. As the number of relevant passages per article goes up, the BEP is placed more toward the start of the article and at higher structural levels, shifting from *context BEP* to *container BEP*.

Table 4: BEP versus FRC over number of relevant passages

# relevant passages	BEP before FRC (%)	BEP at FRC (%)	BEP after FRC (%)
1	634 (17.16)	2,497 (67.60)	563 (15.24)
2	453 (45.21)	354 (35.33)	195 (19.46)
3	229 (60.90)	76 (20.21)	71 (18.88)
4	93 (65.03)	21 (14.69)	29 (20.28)
5	77 (73.33)	14 (13.33)	14 (13.33)
6	32 (78.05)	5 (12.20)	4 (9.76)
7	29 (72.50)	7 (17.50)	4 (10.00)
8	18 (69.23)	4 (15.38)	4 (15.38)
9	7 (87.50)	0 (0.00)	1 (12.50)
10+	38 (82.60)	4 (8.70)	4 (8.70)

3. CONCLUSIONS

We investigated the relation between the best entry point of articles and the relevant information contained in them. Our analysis of the INEX 2006 assessment data leads to the following findings: First, although Wikipedia articles are short on average, the majority of BEPs are not at the start of the article. This signals a preference for a more focused entry point than the article's beginning. This provides support for focused retrieval such as the XML element retrieval studied at INEX, where an important question is what to return to users as answer to their queries. Second, the majority of BEPs coincide with the first relevant character in the text (*browsing BEPs*). This suggests that judging relevant information in documents, as is practiced at INEX, has a meaningful relation to what is regarded as a starting point for accessing relevant information. Third, whereas the majority of BEPs in case of a single relevant passage are *browsing BEPs*, in case of multiple relevant passages the majority of BEPs are *context BEPs*, i.e., objects appearing before the relevant objects, but at the same structural level, or *container BEPs*, i.e., objects which contain several relevant objects.

Our findings also highlight that the relation between best entry point and topically relevant information is not perfect. This is in line with the earlier study of Shakespeare plays [5]. In structured information retrieval, relevance judgments play a broader role than in standard test collections. Apart from locating where the relevant information can be found, they also determine what systems should return to users. Our findings suggest that this depends on more than topical relevance alone, and should take the document structure and distribution of relevant information inside the article into account.

Acknowledgments INEX is an activity of the Network of Excellence in Digital Libraries, DELOS. 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] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 40:64–69, 2006.

[2] K. Finesilver and J. Reid. User behaviour in the context of structured documents. In *Proceedings ECIR 2003*, volume 2633 of *LNCS*, pages 104–119. Springer, 2003.

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

[4] G. Kazai, M. Lalmas, and J. Reid. Construction of a test collection for the focussed retrieval of structured documents. In *Proceedings ECIR 2003*, volume 2633 of *LNCS*, pages 88–103. Springer, 2003.

[5] J. Reid, M. Lalmas, K. Finesilver, and M. Hertzum. Best entry points for structured document retrieval: Parts I & II. *Information Processing and Management*, 42:74–105, 2006.