On the Relation between Relevant Passages and XML Document Structure

Jaap Kamps, Marijn Koolen
University of Amsterdam

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

Motivation

- As Salton et al. [1993, 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.

- But what to return?
  - ⋆ parts of the document structure? or fixed window passages?

Outline

- How does relevant text inside a document relate to the document structure?
- Three questions:
  - ⋆ What is the length of relevant passages? What fraction 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?
- Conclusions

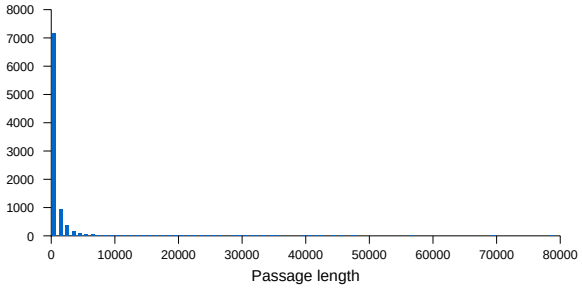
INEX 2006 test collection

- Documents
  - ⋆ English Wikipedia pages transformed into XML
- Topics
  - ⋆ 114 topics created by INEX 2006 participants
- Assessments
  - ⋆ Peer assessments on pooled set of articles
  - ⋆ Assessors mark in yellow all and only relevant text.
  - ⋆ Judges view the rendered text (not the precise XML structure)

Relevant Passage Length

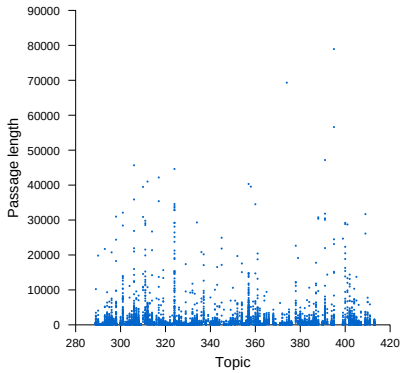
- First we look at the resulting highlighted passages
  - ⋆ What is the length of relevant passages? What fraction of the article is considered relevant?
- Over all topics
  - ⋆ there are 9,086 passages
  - ⋆ in 5,648 articles
- So 1.6 relevant passage per article with relevance.

Relevant Passage Length Distribution



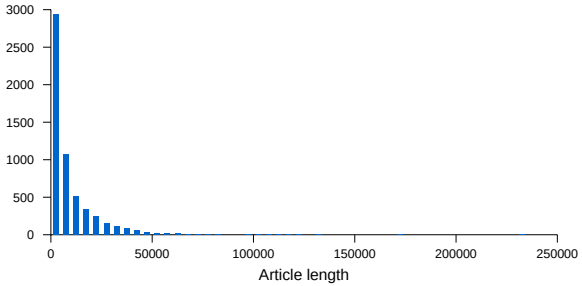
- Passages are short:
  - ⋆ mean length of 1,090 characters (paragraph)
  - ⋆ median length of 297 characters (couple of sentences)

Relevant Passage Length per Topic



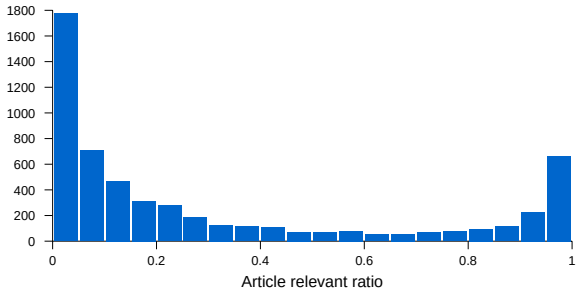
- No (clear) topic effect
  - ⋆ Certainly no "fixed" passage length

Article Length Distribution



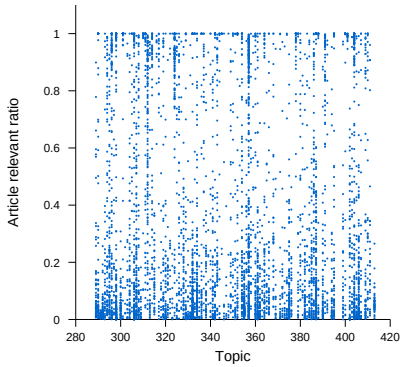
- Length of articles (relevant for any topic):
  - ⋆ Mean length 9,485 characters, median length 4,528

Article Relevant Ratio Distribution



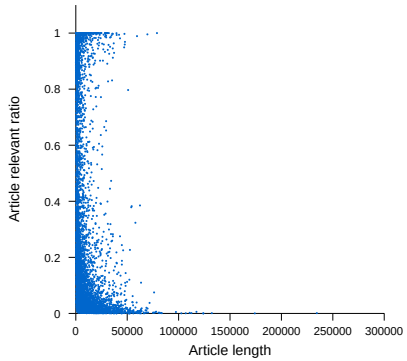
- Coverage spread over the whole range
  - ⋆ Mean ratio is 0.3160, median ratio is 0.1339
  - ⋆ Higher frequently at the extremes.

Article Relevant Ratio per Topic



- There is no (clear) topic/assessor effect
  - ⋆ spread over fractions for all topics

Article Relevant Ratio versus Article Length



- Highlighted ratio does not depend on length
  - ⋆ Contrary to Salton's intuition!
  - ⋆ At least, on Wikipedia...

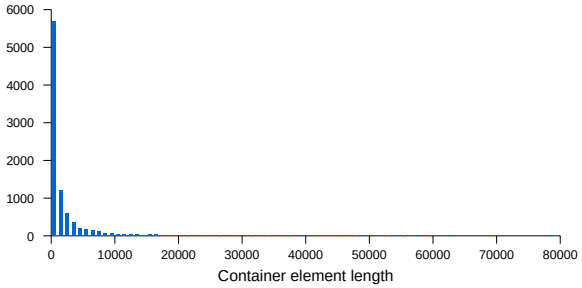
Wrap up: Relevant Passage Length

- We found that
  - ⋆ Relevant passages are relatively short with a median length of a couple of sentences, and an average length of a paragraph
  - ⋆ There is no "fixed" length of relevant passages
  - ⋆ The highlighted text may cover any fraction of the article; and
  - ⋆ The fraction of the article that is highlighted does not depend on the length of the article.

Relating Passages to Elements

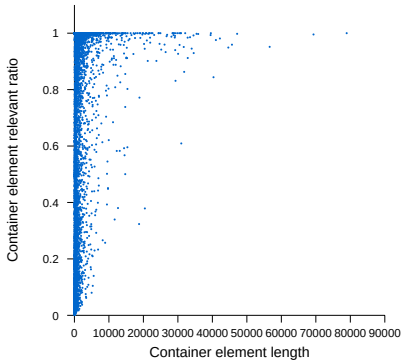
- We'll now look at the document structure
  - ⋆ How well do the highlighted passages correspond to XML elements of the document structure?
- Specifically, we look at
  - ⋆ Container elements = the shortest XML element containing the whole passage

Container Element Lengths



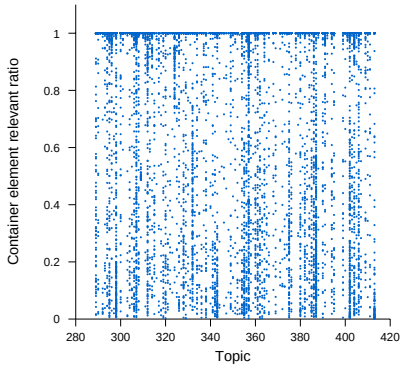
- The smallest element containing the whole passage
  - ⋆ has a mean length of 2,348 and a median length of 620
  - ⋆ that's roughly twice the length of passages!

Container Element Length versus Ratios



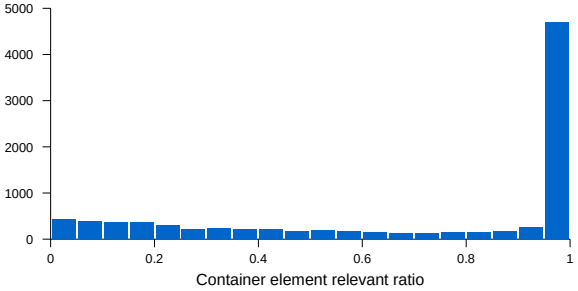
- Highlighted ratio does depend on length container
  - ⋆ Also a result of the definition of container element

Container Element Relevant Ratios per Topic



- Spread over all ratios
  - ⋆ No (clear) topic effect
  - ⋆ Thick line at ratio 1

Container Element Relevant Ratios



- Coverage spread over the whole range
  - ⋆ Mean ratio is 0.7028, median ratio is 0.9730
  - ⋆ 51.8% have 95% or more coverage, 45% have 100%.

Container Element Tags

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

Wrap-up: Relating Passages to Elements

- Mixed results for the correspondence between relevant passages and container elements:
- 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).

Passage and Element Boundaries

- Highlighted passages may span a range of elements
  - ⋆ How do the passage boundaries correspond to XML element boundaries?
- We'll look at:
  - ⋆ start element = the XML element that directly contains the first highlighted character
  - ⋆ end element = the XML element that directly contains the last highlighted character

Start/End Element Distances

	Min	Max	Median	Mean	Stdev
start element offset	0	10,723	0	62.74	317.68
end element offset	0	61,743	2	365.80	2,423.29

- We see that the distance between passage and element boundary
  - ⋆ is at the start 63 (mean) and 0 (median)
  - ⋆ and at the end 367 (mean) and 2 (median)
- Recall passage length is 1,090 (mean) and 297 (median)
- So highlighted passages
  - ⋆ start very near an element boundary
  - ⋆ end close to an element boundary

Container Element Distances

- Let first look at the container elements

	Min	Max	Median	Mean	Stdev
start container offset	0	47,510	1	252.90	1,344.91
end container offset	0	68,566	24	1,023.48	3,928.68

- We see that the distance between container and passage
  - ⋆ is at the start 253 (mean) and 1 (median)
  - ⋆ and at the end 1,023 (mean) and 24 (median)
- That is
  - ⋆ passages tend to start at the container element's start
  - ⋆ but passages end somewhat earlier than the container element

Conclusions: Relevant Text versus Document Structure

- We analyzed the INEX 2006 assessments to investigate how relevant text inside a document relates to the document structure.
- Relevant passages
  What is the length of relevant passages?
  What fraction of the article is considered relevant?
  - ⋆ passages are short: median a few sentences, mean a paragraph
  - ⋆ no "fixed" passage length, but great length variation
  - ⋆ % of the article highlighted also varies greatly, also over topics
  - ⋆ no relation with article length (on Wikipedia)!

Relevant Text versus Document Structure (cont'd)

- Passages versus elements
  How well do the highlighted passages correspond to XML elements of the document structure?
  - ⋆ average (mean) container element length is twice average (mean) passage length
  - ⋆ half of the containers fit passages well (95% or more)
  - ⋆ best fitting document structure: paragraphs, sections, list-items, titles, and the whole article.
- Passage boundaries versus element boundaries
  How do the passage boundaries correspond to XML element boundaries?
  - ⋆ Passage start is an element boundary, the end is close to one
  - ⋆ Passages start at their container element, but end earlier

But What To Retrieve?

- Mixed support for element retrieval and for passage retrieval:
  - ⋆ the short length of the typical relevant passage suggests retrieving fixed window passages,
  - ⋆ but the variation in length of passages and coverage of the article suggests a flexible unit of retrieval like XML elements.
  - ⋆ 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.
- Start of a relevant passage coincides with start of an element
  - ⋆ if we assume results are displayed in the context of the article, retrieval of XML elements seems a good approach

Take Home Messages

- Caveats:
  - ⋆ "container elements" may be very unpredictable...
  - ⋆ finding exact passages is not the INEX task
- What to retrieve?
  - ⋆ Elements seems attractive, esp. sectioning structure
- Recall: relevant text usually the initial part of good elements
  - ⋆ Element trimming (tag/length specific) should be successful
- What does it mean to retrieve an element?
  - ⋆ Users presumably start reading at the beginning
  - ⋆ How bad is it to return trailing non-relevant text?

References

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.