

Passage Retrieval and other XML-Retrieval Tasks

Andrew Trotman (Otago)

Shlomo Geva (QUT)

Passage Retrieval

Information Retrieval

- Information retrieval (IR) is the science of searching for information in documents, searching for documents themselves, searching for metadata which describe documents, or searching within databases, whether relational stand-alone databases or hypertext networked databases such as the Internet or intranets, for text, sound, images or data. (Wikipedia, 2006)

XML-Retrieval

- Information Retrieval from document collections in which at least one component is marked up in XML
- What is a document?
 - Yes, well, lets not go there
- What is a component?
 - Lets not go there either

XML Element Retrieval

- Aim:
 - To identify a more focused result to a query than a whole document, that is, to identify a relevant ***fragment*** of a document
- This involves identifying two factors:
 - The ***location*** of a relevant fragment
 - The ***size*** of the relevant fragment
- Where the fragment is an XML element

An XML Document

```
<?xml version="1.0" encoding="UTF-8"?>
<article>
  <name id="59186">Dreamland, Michigan</name>
  <conversionwarning>0</conversionwarning>
  <body>
    <redirectlink src="Torch Lake Township,
                      Houghton County, Michigan">
      Torch Lake Township, Houghton County, Michigan
    </redirectlink>
  </body>
</article>
```

Why XML Element Retrieval?

- XML is used to identify semantic elements
- The user's information need is semantic
- Given:
 - XML is used correctly
 - The query is sensible
- Reasonable Assumption:
 - The best fragments are whole XML elements
- The ***task*** is to identify these elements

Really?

- Are XML Elements really the best fragments?
- We draw our conclusion from other studies...

Agreement Levels

- Pehcevski, Thom & Vercoustre (2005)
 - Topics used in INEX 2004 interactive experiments
 - Agreement only at the ends of the relevance scale
 - E0S0 and E3S3
 - This is also true of the Cystic Fibrosis collection
 - Conclusion
 - Obviously relevant and obviously not is obvious
 - Levels of relevance are debatable
 - The 10 point relevance scale isn't needed

Agreement Levels

- Trotman (2005)
 - 12 topics double judged at INEX 2004
- Pehcevski & Thom (2005)
 - 5 topics double judged at INEX 2005

Evaluation	Agreement ($\cap/\cup$)
TREC 4 P/B	0.49
TREC 4 A/B	0.43
TREC 4 P/A	0.42
TREC 6	0.33
INEX 2004 documents	0.27
INEX 2004 elements	0.16
INEX 2005 documents	0.39
INEX 2005 elements	0.24

Please, someone,
Kappa these

Passages and Elements

- At INEX 2005 judges
 - Were presented with pool documents
 - Marked relevant passages in those documents
 - Set exhaustivity values to elements in the passages
 - The exhaustivity of an element was this score
 - The specificity was the proportion of text highlighted
- Conclusions
 - Judging passages is easier than judging elements

Which Elements are Relevant?

- Trotman & Lalmas (2006)
 - INEX 2005 judgments
 - Most relevant elements were paragraphs
 - Regardless of query (target element)
 - Most relevant elements were paragraphs
 - Regardless of thorough / focused
 - Most relevant elements were paragraphs
 - Conclusion
 - Assessors are identifying relevant sequences of paragraphs

How Relevant?

- Piwowarski, Trotman & Lalmas (2006)
 - INEX 2005 judgments
 - Examined average specificity of element and found:

Element	Average Specificity
Paragraph	0.94
Section	0.51
Body	0.15
Article	0.12

- Conclusion
 - Paragraphs are either relevant or not
 - Judges are identifying collections of paragraphs

Elemental Passages

- Piwowski, Trotman & Lalmas (2006)
 - INEX 2005 judgments
 - Elemental passages
 - A passage that is an element
 - Starts and ends on the boundaries of a single element
 - Found:
 - 36% of passages were elemental
 - 64% of passages were not elemental
 - Conclusion
 - Judges identified relevant passages, not relevant elements

Elements and Assessment

- Ogilvie & Lalmas (2006)
 - INEX 2005 judgments
 - Examine the stability of metrics using just specificity
 - Discover
 - Remove exhaustivity from the assessments and...
 - The relative performance of search engines remains stable
 - Conclusion
 - Using specificity alone is sufficient for assessment purpose
 - Specificity is based on highlighting passages not elements

Passage Agreement?

- INEX 2005
 - Piwowarski, Trotman & Lalmas (2006)
 - Passage agreement
 - Pehcevski & Thom (2005)
 - Document and element agreement

Evaluation	Agreement ($\cap/\cup$)
Documents (binary)	0.39
Elements (exact)	0.24
Passages	0.23

XML Passage Retrieval

- The case for passage retrieval is compelling:
 - Element agreement level is higher with highlighting
 - The relevant text is a collection of paragraphs
 - Not an element
 - Assessment is stable with highlighting
 - Passage agreement levels look fine
- Some problems vanish:
 - “too small” elements can’t occur
 - Conversion of highlights to elements isn’t needed

Three New Tasks

- Focused[†] Retrieval
 - The identification of non-overlapping passages of text relevant to the user's information need
 - Sorted by passage relevance
- Relevant in Context
 - The identification of non-overlapping passages of text relevant to the user's information need
 - Sorted by document and sequential within each document
- Best in Context
 - The identification of the best entry point (BEP) within a document
 - Sorted by document

[†] one 's' this time

Transitioning to Passages

- The transition from elements to passages can gradual
- To convert from an element to a passage is straightforward
 - Use the element as a passage
 - Possibly merge adjacent passages
 - Clarke (2005)
 - XML range specification

Passages at TREC

- TREC HARD
 - 2003 and 2004
- TREC Genomics
 - 2006
- Perhaps we should be sharing resources
 - Same documents different formants?
- Metrics?
 - TREC and INEX both already have passage metrics

Other XML-Retrieval Tasks

The Performance Task

- What are the best ranking algorithms for:
 - The web?
 - Hits?
 - PageRank?
 - Whatever Google actually uses?
 - Unstructured text?
 - BM25?
 - Pivoted length-normalized Cosine?
 - Language Models?
 - XML?
 - Suggestions from the floor please... (focused or thorough)

Five Years...

- How do we define best?
 - But the metrics change from year to year!
- Work to be done
 - Identify the current state of the art
 - Change methodology so we compare to that
 - Perform statistical tests
- Homework...
 - Trawl through old runs with new metrics
- Future...
 - Can we find a way to measure incremental improvements?
 - Withhold some judgments for re-use in future years?

Multiple Document Formats

- The premise is that XML is helpful in retrieval
- Experiment on different granularities of XML might be performed to show this. Does XML actually help?
- Either reduce the density of tags or use the same document collection in different formats (XML / HTML / TEXT)

Related Articles

- Inserting and maintaining links in web pages and the Wikipedia is time-consuming and tedious
- Can we identify methods of predicting which pages in a closed set of pages should link to each other?
- The Wikipedia offers a unique environment
 - The pages are already cross linked
 - We can measure performance by similarity to these
 - Remove them, then try and predict them
 - Topics would be Wikipedia documents chosen at random

Question Answering

- The Wikipedia is an obvious collection to do question answering
- XML element might directly contain the answers to questions (Wikipedia templates)
- XML elements might be used to identify relevant parts of documents from which answers are extracted

Conclusions

- We see that
 - Passage retrieval is well suited to XML
 - Element retrieval is not well suited to XML
 - Passage tasks are easy to specify
 - Passages are easier to judge
- So
 - XML retrieval should focus on passages not elements
- And
 - There's plenty of other tasks too