Document & Passage Retrieval

Chase Hermsen, Sergei Lushtak, Joshua Lutes

Contents

Initial Query Reformulation

Queries in series often don't contain enough information in isolation:

When was it formed?
When was he born?

Use an implementation of some of Shaw et al. 08's features for reformulating queries greatly improves standaloneness of questions.

Example from Diff of Reformulation

< 59.2 When was it formed?
< 59.3 What is its purpose?
< 59.4 How many members does it have?
< 59.5 Who is its current head?
< 60.1 What state does he represent?
< 60.2 When was he born?
< 60.3 When was he first elected to the senate?
< 60.4 What branch of the service did he serve in?

> 59.2 When was Public Citizen formed ?
> 59.3 What is Public Citizen 's purpose ?
> 59.4 How many members does Public Citizen have ?
> 59.5 Who is Public Citizen 's current head ?
> 60.1 What state does senator Jim Inhofe represent ?
> 60.2 When was senator Jim Inhofe born ?
> 60.3 When was senator Jim Inhofe first elected to the senate ?
> 60.4 What branch of the service did senator Jim Inhofe serve in ?

Ngram document reranking

- Rerank documents based on ngram occurrence and frequency
- Weighted according to
  - Size of ngram
  - Frequency of ngram

Training set: Test2003.txt
Local Analysis

Add top 10 termrs and 50 bigrams w/o stop words from to 5 - 100 documents. Heavily reweight the original query.

Improvement on some queries but over all worse than the baseline.

Training set: Test2003.txt
Local Context Analysis

Add top 20 concepts (noun sequences) from top 3 documents. Slightly reweight the original query.

Improvement on some queries but over all worse that the baseline.

Problems: Tuning on a small number of queries is not reliable. Using large number of queries is difficult; run can take hours.

Passage Retrieval

- Standard passage retrieval using Lemur and Indri
  - Passage size of 200, every 100 characters
- Problem of pulling documents with frequent use of the terms in the query but no answer to the question.
  - *Lady complaining about time zones, time zones, time zones*
- Problem of the passage being retrieved from the "wrong" part of the document
  - *&QL; I love living in Cambridge for so many reasons, many of them wrong, I suppose, according to the Chamber of Commerce. But I love it just the same. It's an incredibly tolerant city, as long as you'*
- Problem of multiple duplicate passages being retrieved from different documents

Information Retrieval

0.1585 baseline 5
0.2654 chase 5

0.1230 baseline 4
0.2957 chase 4

Passage Retrieval Evaluation & Improvements

- MRR on TREC-2005: 0.0670
- Documents should probably be cleaned up
  - Removing non-textual information seems beneficial with no bad effects
  - Getting rid of duplicate documents
  - This is hard since we done gots to copy the corpus over
- Without cleaning up the documents we should still be able to do something with duplicate passages, push them down to the bottom when reranking, perhaps?