Search Engines

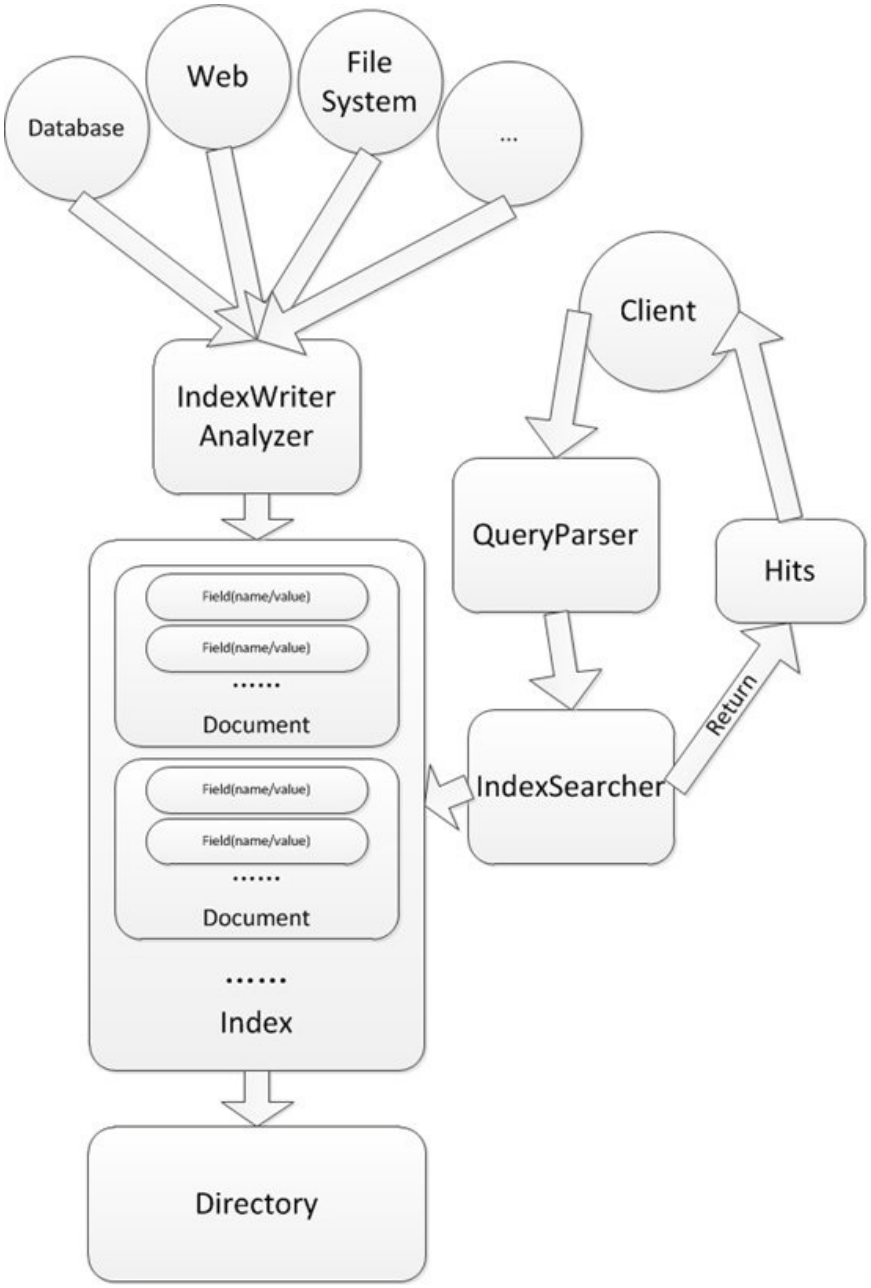
Exercise 5: Querying

Dustin Lange & Saeedeh Momtazi

9 June 2011

Task 1: Indexing with Lucene

- We want to build a small search engine for movies
- Index and query the titles of the 100 best rated movies of all time from IMDb with Lucene



Source: http://blogs.msdn.com/b/windows-azure-support/archive/2010/11/01/how-to-use-lucene-net-in-windows-azure.aspx

Lucene Components

Indexing

- Directory
  - Where to put the index: RAM / file system
- Document
  - Entity to be indexed / searched
  - Consists of fields
- Analyzer
  - Prepare indexed text
  - Tokenization, stemming, stop words, …
- IndexWriter
  - Create and maintain index

Searching

- QueryParser
  - Parse query, return query object
- IndexSearcher
  - Access index with query object
- Collector
  - Scoring results (Hits) and result filtering

Task 1: Indexing with Lucene

- Download Apache Lucene 3.1.0
  - http://lucene.apache.org/java/docs/index.html
  - http://www.apache.org/dyn/closer.cgi/lucene/java/
- Compile and run the given Java class
- Change the implementation so that
  a) Stop words are removed
  b) Stemming is done
  c) Suggestions for single-term queries are made (based on the indexed data)
  d) Suggestions for phrase queries are made (based on the indexed data)
- Which changes did you make? Briefly explain how your solution works (i.e., how Lucene's components work).
- How do the results of the given queries improve with each of these changes? (For the suggestions, compare the results of the original query with the results of the best suggestion.)
- You will need *lucene-core*, but also additional packages from the *contrib* folder (*analyzers, spellchecker*)

Task 2: Positional Pseudo-relevance Feedback

- Relevance feedback
  - User identifies relevant documents in the initial result list
  - Modifying query using terms from those documents and re-ranks documents
- Pseudo-relevance feedback
  - System assumes all top retrieved documents for initial query are relevant
  - Expanding query based on the words that are co-occurred with query terms in top documents
- Localized/Positional pseudo-relevance feedback
  - Intuition: words closer to query words are more likely to be related to the query topic
  - Selecting from feedback documents those words that are focused on the query topic based on positions of terms in feedback documents
  - Exploiting term positions and proximity to assign more weights to words closer to query words
  - Co-occurrence counts are measured in a small window of terms around the query words

Task 2: Positional Pseudo-relevance Feedback

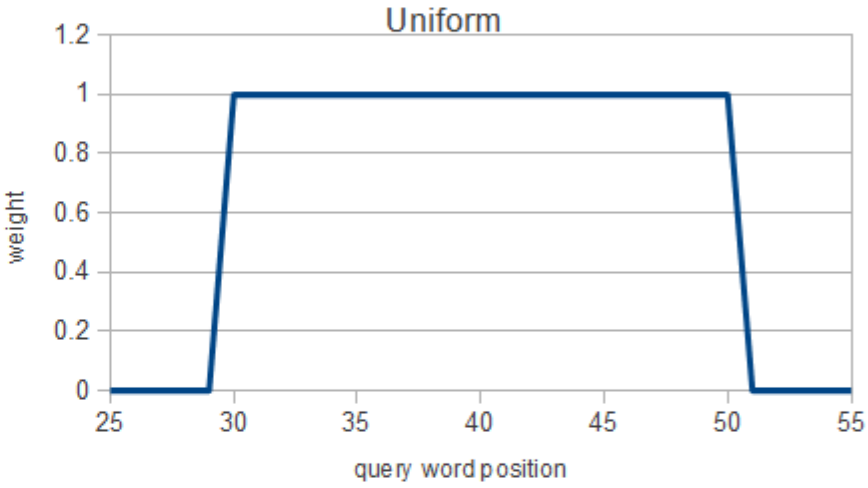
- Use one of the following words as a query in “google.com” and assume the top 20 documents are relevant:
  - Car
  - Camel
  - Golf
- Find top 15 words that are related to the query based on the positional pseudo-relevance model
- Compare the list of related words that are extracted based on the proposed variations (next slide)

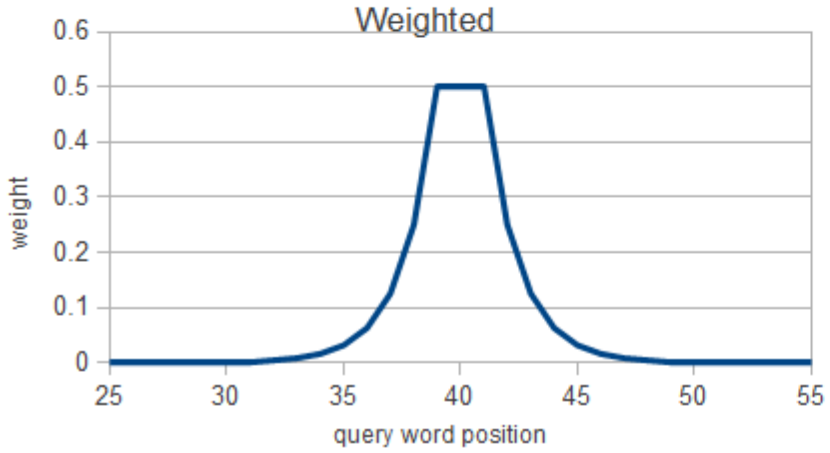
Task 2: Positional Pseudo-relevance Feedback

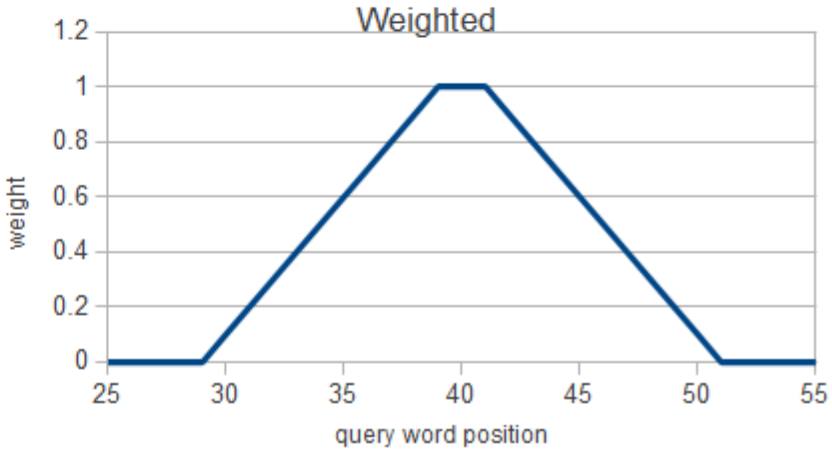
- Use window size 10 with the following term association measures:
  - Dice's coefficient
  - Mutual information
  - Pearson's Chi-squared ($\chi 2$) measure
- Use window size 10 with the following localized measures (see next slide for examples):
  - Uniform
  - Weighted
- Use uniform model and mutual information with the following contexts:
  - Whole document
  - 5 words

Task 2: Positional Pseudo-relevance Feedback

- Localized measures
  - Uniform
  - Weighted







Submissions & Next Exercise

- Submissions:
  - Create slides to present your solution.
  - Send your presentation
    - as PDF or PPT(X) or ODP: *SearchEngines**5**[Name1][Name2].[pdf|ppt|pptx|odp]*
    - via e-mail with subject: *Search Engines **5***
    - to *dustin (dot) lange (at) hpi (dot) …*
    - until **04 July 2011, 5:00 pm**
- On **05 July 2011 (Tuesday)**: Be prepared to present your solution
  - English (or German)
  - Absent: Send us an e-mail in advance

Thanks for Listening

- Updates
  - Mailing list: searchengines2011 (at) hpi (…)
  - See website
- Questions
  - Via e-mail:
    - dustin (dot) lange (at) hpi (…)
    - saeedeh (dot) momtazi (at) hpi (…)
  - Office: A-1.6 / A-1.7