Practical Relevance

Grant Ingersoll

Thinking Lucene ◥ Think Lucid.

Two Tales of Relevance

- **The case of the missing data**
- The power of suggestion

Agenda

- Why Tune Relevance?
- Look before leaping
- Tips, traps and techniques for improving relevance

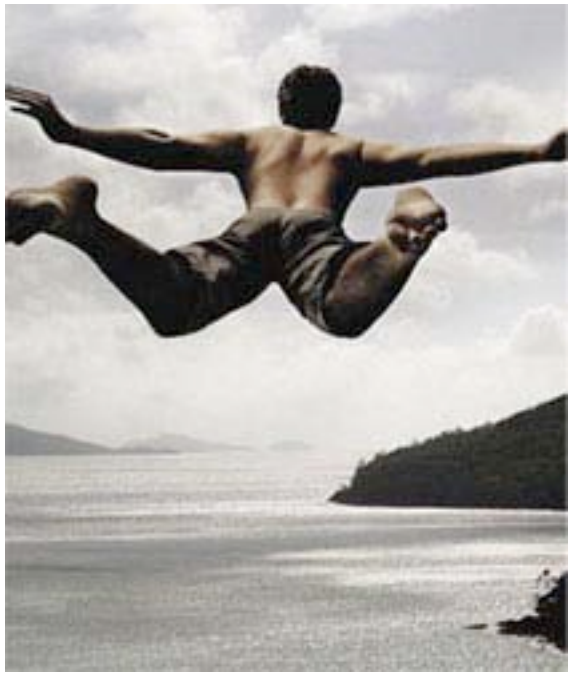
Why Tune Relevance?

- Better search results = Less time searching, more time acting
- Less time searching = Happier, more effective users
- Happier, more effective users = $, €, £ (earned/saved)
- $, €, £ (earned/saved) = Big fat raise for you!

What does “better search” mean to you?

Look Before You Leap

- Before undertaking any relevance tuning, you need to define what "better search" means to you
- Once determined, many ways to test/measure
- Once tested, many ways to fix



http://www.betternetworker.com/files/useruploads/16675/leap.jpg

Testing Relevance

- A/B testing
- Log Analysis
- Empirical
  - Top X queries, plus random sample
- Ask
  - Ratings/Reviews
  - Focus Groups
- Also: Ad Hoc, Open Relevance Proj. TREC, etc.

http://www.flickr.com/photos/spaceamoeba/22351055/

Tips, Traps and Techniques

- High Level and abstract concepts that are agnostic of search engine
- Lucene/Solr specific
- Focus on “big/easy wins”

Understand your…

- **Domain**
  - Types of documents
  - Languages present
  - Document structures, metadata and other features
  - Lexical resources: jargon, synonyms, abbreviations...
  - Relationships between documents
- **Users**
  - Sophistication/Expertise
  - Search and Discovery needs
  - Known Item vs. Keyword
- **Tolerance for Pain**
  - Managers
  - Business Interests
  - Release cycles
  - Obsession in finding the one true relevance model (hint, it doesn't exist)
  - "explain() blindness"

High Level

- **Leverage a priori information for ranking**
  - Link Authority (and other graph ranking approaches)
  - User ratings/reviews
  - Timestamps
  - Classification
  - NLP
- **Gather and leverage user feedback**
  - Relevance feedback (manual and psuedo)
  - Log analysis (click analysis)
  - Ratings/reviews
  - Filters and Facets

Debugging

- Check the analysis (more in the next slide)
- Check for data quality issues
- Check your query constructs (slops, boosts, etc.)
  - Try alternate query representations
  - AND vs. OR
- Use Lucene's explain() or Solr's &debugQuery

Analysis

- **More often than not, when a query doesn't match, there is an analysis problem**
- **Debugging Analysis:**
  - Luke -- http://code.google.com/p/luke/
  - Solr built-in: http://localhost:8983/solr/admin/analysis.jsp
- **Common reasons for mismatch:**
  - Stemming, wrong case, compound word, spelling/fuzzy (iPod, i Pod, i-Pod)

Phrases

- Almost always a win to automatically add phrase query variations to all multiword queries
  - Even better to detect key phrases, but…
- In Solr, with the (e)Dismax handler, use the &pf and &ps options to automatically add phrase boosts
- Use stopwords with phrases
- Using a large slop factor can simulate an AND query while rewarding close proximity
- See also the ComplexPhraseQuery in contrib/queryparser
- Consider SpanQuery and derivatives

Boosts

- **Index time boosts to consider:**
  - Document and field boosts -- only offer a small bit of precision – tread carefully
  - Payloads can be used to boost individual terms
  - ExternalFileField (Function Query) can be used for higher precision document boosts
- **Search Time boosts to consider:**
  - Term, Field, Recency and Other Functions
- **Function Queries (in Solr) are your friend**
- **(e)Dismax**

Discovery Tools

- Facets
- Filters
- Auto-suggest
- Spellchecking (Did you mean?)
- Related Searches
- Related Items

Deep Relevance – Here be Dragons

- **Write your own Query/Weight/Scorer**
- **Customize Lucene's Similarity**
  - Lucene tends to favor shorter documents by default
- **Flexible Indexing opens up a whole new opportunity for alternate scoring models**
  - BM25, Language Modeling, others
  - Type-safe payloads
  - ???
- **Natural Language Processing for semantic information**

Resources

- ACM SIGIR - http://sigir.org/
- http://lucene.li/X
- http://lucene.li/Z
- Open Relevance Project: http://lucene.apache.org/openrelevance
- grant@lucidimagination.com
- http://www.lucidimagination.com
- @gsingers