Title

Static index pruning

Lucene MeetUp Lightning Talk

Andrzej Bialecki
<ab@getopt.org>

SIGRAM

Index composition

- Stored fields
- Terms dictionary
- Term frequency data
- Positional data (postings)
  - With or without payload data
- Term frequency vectors

Top-N retrieval

- N is usually << 1000
- Number of documents may be into millions
- Number of terms usually is well into millions
- Not to mention individual postings …

- Question: do we really need to keep ALL of this information for a good-quality top-N search for common queries?

Pruning hypothesis

- There should be a way to remove some of the less important data
  - While retaining the quality of top-N results!
- Question: what data is less important?
- Some answers:
  - That of poorly-scoring documents
  - That of common (less selective) terms
- Can we do this work in advance (static pruning)?

Stored field pruning

- Some stored data can be compacted, removed, or restructured:
- Use case: generating “snippets”
  - Split content into sentences
  - Reorder sentences by a static “importance” score (e.g. how many rare terms they contain)
  - Remove the bottom x% of sentences

What do we need for top-N ?

- Work backwards
- “Foreach” common query:
  - Run it against the full index
  - Record the top-N matching documents
  - “Foreach” document:
    - Record the terms and term positions that contributed to the score
- Finally: remove all non-recorded postings and terms
- First proposed by D. Carmel (2001) for single term queries

Smarter pruning

Before pruning
0 quick
1 brown
2 fox

After pruning
0 quick
1 brown
2 fox

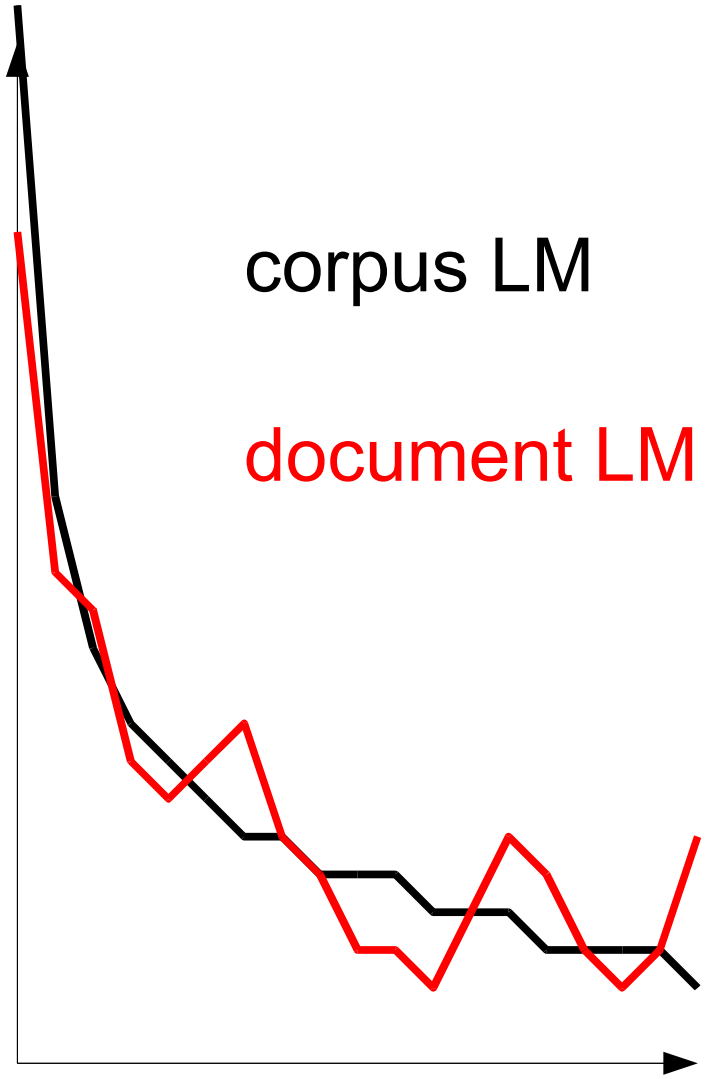
Query 1: brown - topN(full) == topN(pruned)

Query 2: "brown fox" - topN(full) != topN(pruned)

- Hmm, what about less common queries?
  - 80/20 rule of "good enough"?
  - Term-level is too primitive
- Document-centric pruning
- Impact-centric pruning
- Position-centric pruning

Metrics of term/posting importance

- Plain in-document term frequency (TF)
- Residual IDF – measure of term informativeness (selectivity)
- Kullback-Leibler divergence from a language model →
- TF-IDF score obtained from top-N results of TermQuery (Carmel method)



Applications of pruned indexes

- Obviously, performance-related
  - Some papers claim a modest impact on quality when pruning up to 60% of postings
- Removal of (some) stored content
  - Legacy indexes, or ones created with a fossilized external chain
- Multi-tier searching:
  - Most common queries handled by a heavily pruned 1st tier index that fits wholly in RAM
  - Less common queries handled by a 2nd tier index (on SSD?)
  - Remaining queries handled by a full index

StorePruningPolicy impl.

- May remove (some) fields from (some) documents
- May as well modify the values
- May rename / add fields

TermPruningPolicy impl.

- Thresholds (in the order of precedence):
  - Per term
  - Per field
  - Default
- Plain TF pruning – TFTermPruningPolicy
  - Removes all postings for a term where TF (in-document term frequency) is below a threshold
- Top-N term-level – CarmelTermPruningPolicy
  - TermQuery search for top-N docs
  - Removes all postings for a term outside the top-N docs

Experiences so far

- Mixed … ☺
- TF pruning: recall of phrase queries becomes very poor
- Carmel pruning – slightly better, but still heavy negative impact on phrase queries
- I'm experimenting now with Kullback-Leibler divergence metric and Residual IDF
- Recognizing and keeping key phrases would help
- Use query log for frequent-phrase mining?
- NEED A TEST FRAMEWORK, to formalize the impact of various pruning methods on P@10, MRR ...

Questions?

- JIRA issue:
  http://issues.apache.org/jira/browse/LUCENE-1812
- References:
  - Static index pruning for information retrieval systems, D. Carmel at al, ACM SIGIR 2001
  - Pruned query evaluation using pre-computed impacts, V. Anh et al, ACM SIGIR 2006
  - A document-centric approach to static index pruning in text retrieval systems, S. Buettcher et al, ACM SIGIR 2006
  - Pruning Policies for Two-Tiered Inverted Index with Correctness Guarantee, A.Ntoulas et al, ACM SIGIR 2007
  - A Practitioner's Guide for Static Index Pruning, I. Altingovde et al, ECIR 2009
- Contact the author: ab@getopt.org