Okapi BM25

- Problem: BIM favors long documents
  - More likely to contain matching terms
- Solution:
  - Inter-document frequency for “relevance”
  - Scale by document length
  - Scale by query length/term frequency
Okapi BM25

- BM25 metric, used in Okapi IR system
  - $V$=relevant documents, $VNR$=not relevant

\[
RSV_d = \sum_{t \in q} \left[ \frac{|VR_t| + \frac{1}{2}}{|VNR_t| + \frac{1}{2}} \right] \times \frac{(k_1+1)tf_{td}}{k_1 \left(1-b\left(\frac{L_d}{L_{ave}}\right)+b\right) + tf_{td}} \times \frac{(k_3+1)tf_{sq}}{k_3 + tf_{sq}}
\]

Probabilistic interpretation of IDF

Term frequency with normalization by length

Query term weighting

PRP and BIM

- Getting reasonable approximations of probabilities is possible.

- Requires restrictive assumptions:
  - Term independence
  - Terms not in query don’t affect the outcome
  - Boolean representation of documents/queries/relevance
  - Document relevance values are independent

- Some of these assumptions can be removed
- Problem: either require partial relevance information or only can derive somewhat inferior term weights
Removing term independence

- In general, index terms aren’t independent
- Dependencies can be complex
- van Rijsbergen (1979) proposed model of simple tree dependencies
  - Exactly Friedman and Goldszmidt’s Tree Augmented Naive Bayes (AAAI 13, 1996)
- Each term dependent on one other
- In 1970s, estimation problems held back success of this model

And to put some faces to the names you’ve been seeing…

Karen Spärck Jones
Stephen Robertson
Keith van Rijsbergen
CS47300: Web Information Search and Management

Web Search
Prof. Chris Clifton
10 September 2019
Some slides courtesy
Manning, Raghavan, and Schütze
Without search engines the web wouldn’t scale

- No incentive in creating content unless it can be easily found – other finding methods haven’t kept pace (taxonomies, bookmarks, etc)
- The web is both a technology artifact and a social environment
  - “The Web has become the “new normal” in the American way of life; those who don’t go online constitute an ever-shrinking minority.” – [Pew Foundation report, January 2005]
- Search engines make aggregation of interest possible:
  - Create incentives for very specialized niche players
    - Economical – specialized stores, providers, etc
    - Social – narrow interests, specialized communities, etc

Without search engines the web wouldn’t scale

- The acceptance of search interaction makes “unlimited selection” stores possible:
  - Amazon, Netflix, etc
- Search has been the best mechanism for advertising on the web, a $15+ B industry.
  - Growing very fast but entire US advertising industry $250B – huge room to grow
  - Sponsored search marketing is about $10B
Relevance
- For each query Q and stored document D in a given corpus assume there exists relevance $\text{Score}(Q, D)$
  - Score is average over users $U$ and contexts $C$
- Optimize $\text{Score}(Q, D)$ as opposed to $\text{Score}(Q, D, U, C)$
- That is, usually:
  - Context ignored
  - Individuals ignored
  - Corpus predetermined

Bad assumptions in the web context

The coarse-level dynamics

Content creators

Content aggregators

Content consumers

Editorial

Subscription

Advertisement

Crawls

Feeds
Brief (non-technical) history

Early keyword-based engines
– Altavista, Excite, Infoseek, Inktomi, ca. 1995-1997

Paid placement ranking: Goto.com (morphed into Overture.com → Yahoo!)
– Your search ranking depended on how much you paid
– Auction for keywords: *casino* was expensive!

1998+: Link-based ranking pioneered by Google
– Blew away all early engines: Great user experience in search of a business model
– Meanwhile Goto/Overture’s annual revenues were nearing $1 billion

Result: Google added paid-placement “ads” to the side, independent of search results
– Yahoo follows suit, acquiring Overture (for paid placement) and Inktomi (for search)
Google has maintained that ads (based on vendors bidding for keywords) do not affect vendors’ rankings in search results.

**Ads vs. search results**

Other vendors (Yahoo, MSN) have made similar statements from time to time

– Any of them can change anytime

We will focus primarily on search results independent of paid placement ads

– Although the latter is a fascinating technical subject in itself
Web search basics

User Needs

<table>
<thead>
<tr>
<th>Need [Brod02, RL04]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Informational</strong> – want to learn about something</td>
</tr>
<tr>
<td>(~40% / 65%)</td>
</tr>
<tr>
<td><strong>Navigational</strong> – want to go to that page (~25% / 15%)</td>
</tr>
<tr>
<td><strong>Transactional</strong> – want to do something (web-mediated)</td>
</tr>
<tr>
<td>(~35% / 20%)</td>
</tr>
<tr>
<td>• Access a service</td>
</tr>
<tr>
<td>• Downloads</td>
</tr>
<tr>
<td>• Shop</td>
</tr>
<tr>
<td><strong>Gray areas</strong></td>
</tr>
<tr>
<td>• Find a good hub</td>
</tr>
<tr>
<td>• Exploratory search “see what’s there”</td>
</tr>
</tbody>
</table>
Web search users

- Make ill defined queries
  - Short
    - AV 2001: 2.54 terms avg, 80% < 3 words
    - AV 1998: 2.35 terms avg, 88% < 3 words
    [Silv98]
  - Imprecise terms
  - Sub-optimal syntax (most queries without operator)
  - Low effort
- Wide variance in
  - Needs
  - Expectations
  - Knowledge
  - Bandwidth

- Specific behavior
  - 85% look over one result screen only
  - 78% of queries are not modified (one query/session)
  - Follow links – “the scent of information”

Query Distribution

Power law: few popular (typically broad) queries, many rare (typically more specific) queries
How far do people look for results?

“When you perform a search on a search engine and don’t find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)”

- After reviewing the first few entries: 12%
- After reviewing the first page: 16%
- After reviewing the first 2 pages: 20%
- After reviewing the first 3 pages: 25%
- After reviewing more than 3 pages: 27%

(Source: iprospect.com WhitePaper_2006_SearchEngineUserBehavior.pdf)

Users’ empirical evaluation of results

- Quality of pages varies widely
  - Relevance is not enough
  - Other desirable qualities (non IR!!)
    - Content: Trustworthy, new info, non-duplicates, well maintained,
    - Web readability: display correctly & fast
    - No annoyances: pop-ups, etc
- Precision vs. recall
  - On the web, recall seldom matters
The Web corpus

- No design/co-ordination
- Distributed content creation, linking, democratization of publishing
- Content includes truth, lies, obsolete information, contradictions …
- Unstructured (text, html, …), semi-structured (XML, annotated photos), structured (Databases)…
- Scale much larger than previous text corpora … but corporate records are catching up.
- Content can be dynamically generated

The Web: Dynamic content

- A page without a static html version
  - E.g., current status of flight AA129
  - Current availability of rooms at a hotel
- Usually, assembled at the time of a request from a browser
  - Typically, URL has a ‘?’ character in it
Dynamic content

- Most dynamic content is ignored by web spiders
  - Many reasons including malicious spider traps
- Some dynamic content (news stories from subscriptions) are sometimes delivered as dynamic content
  - Application-specific spidering
- Spiders commonly view web pages just as Lynx (a text browser) would
- Note: even “static” pages are typically assembled on the fly (e.g., headers are common)