Chapter IR:V

V. Retrieval Models

- Overview of Retrieval Models
- Empirical Models
- Boolean Retrieval
- Vector Space Model
- Probabilistic Models
- Binary Independence Model
- Okapi BM25
- Hidden Variable Models
- Latent Semantic Indexing
- Explicit Semantic Analysis
- Generative Models
- Language Models
- Combining Evidence
- Web Search
- Learning to Rank

Web Search

- Retrieval models in practice
  - Web search most important—but not the only—search application
- Major differences to TREC news
  - Size of collection
    - Billions
  - Connections between documents
    - Links
  - Range of document types
  - Importance of spam
  - Volume of queries
    - Tens of billions per day
  - Range of query types
    - Informational, navigational, transactional

Web Search

Search Taxonomy

- Informational
  - Finding information about some topic that may be on one or more web pages
  - Topical search
- Navigational
  - Finding a particular web page that the user has either seen before or is assumed to exist
  - Known-item search
- Transactional
  - Finding a site where a task such as shopping or downloading something can be performed

Web Search

- For effective navigational and transactional search, need to combine features that reflect user relevance
- Commercial web search engines combine evidence from hundreds of features to generate a ranking score for a web page
  - Page content
  - Page metadata
    - “Age,” how often it is updated
    - URL of the page
    - Domain name of its site
    - Amount of text content
  - Anchor text
  - Links (e.g., PageRank)
  - User behavior (click logs)

Web Search

Search Engine Optimization

- SEO: Understanding the relative importance of features used in search and how they can be manipulated to obtain better search rankings for a web page
  - Improve the text used in the title tag
  - Improve the text in heading tags
  - Make sure that the domain name and URL contain important keywords
  - Improve the anchor text and link structure
- Some of these techniques are regarded as not appropriate by search engine companies

Web Search

- In TREC evaluations, most effective features for navigational search are
  - Text in the title, body, headings (h1, h2, h3, and h4)
  - Anchor text of all links pointing to the document
  - PageRank and inlink count
- Given size of Web, many pages will contain all query terms
  - Search engines can use AND semantics
    - Dangerous for smaller collections
      - Site search, news search, ...
      - TREC: Only 50% of relevant pages contain all search terms
  - Ranking algorithm focuses on discriminating between these pages
  - Term proximity is important

Web Search

Term Proximity

- Assumption: Query terms are likely to appear in close proximity within relevant documents
  - `green party political views`
- Many models have been developed
  - n-grams are commonly used in commercial web search
- Dependence model based on inference net has been effective in TREC
  - Let S_q be the set of all non-empty subsets of q (power set)
    - Every $s \in S_q$ that consists of contiguous query terms is likely to appear as an exact phrase in a relevant document
      - Using the `#1` operator in Indri/Galago or `"` in commercial engines
    - Every $s \in S_q$ such that $|s| > 1$ is likely to appear (ordered or unordered) within a reasonably sized window of text in a relevant document
      - Using `#uw8` for $|s| = 2$ and `#uw12` for $|s| = 3$ in Indri/Galago

Web Search

Query types

- Insights gained from TREC experiments
- Topical search:
  - Simple terms and proximity features suffice
- Navigational search:
  - More evidence is helpful
- Pseudo-relevance feedback
  - Helps topical search but detrimental for navigational search
- But: How can we determine query type?

- Other evidence is in general useful
  - User behavior: Clicked-on pages, dwell time, links followed
- But: How to weight and combine more and more evidence?
  - Idea: Machine learning