Information Retrieval

INFO 4300 / CS 4300

- Retrieval models
  - Older models
    - » Boolean retrieval
    - » Vector Space model
  - Probabilistic Models
    - » BM25
    - » **Language models**

Language Model

- *Unigram language model*
  - probability distribution over the words in a language
  - generation of text consists of pulling words out of a "bucket" according to the probability distribution and replacing them
- *N-gram language model*
  - some applications use bigram and trigram language models where probabilities depend on previous words

Language Model

- A *topic* in a document or query can be represented as a language model
  - i.e., words that tend to occur often when discussing a topic will have high probabilities in the corresponding language model
- *Multinomial* distribution over words
  - text is modeled as a finite sequence of words, where there are *t* possible words at each point in the sequence
  - commonly used, but not only possibility
  - doesn't model *burstiness*

LMs for Retrieval

- 3 possibilities:
  - probability of generating the query text from a document language model
  - probability of generating the document text from a query language model
  - comparing the language models representing the query and document topics
- Models of topical relevance

Query-Likelihood Model

- Rank documents by the probability that the query could be generated by the document model (i.e. same topic)
- Start with a query, so calculate P(D|Q) to rank the documents
- Use Bayes' Rule

$$p(D|Q) \stackrel{rank}{=} P(Q|D)P(D)$$

- Assuming prior is uniform, unigram model

$$P(Q|D) = \prod_{i=1}^{n} P(q_i|D)$$

LMs for Retrieval

- 3 possibilities:

QL – **probability of generating the query text from a document language model**

  – probability of generating the document text from a query language model

  – comparing the language models representing the query and document topics

- Models of topical relevance

Query likelihood model

- Simple model
- Directly incorporates term frequency
- Term weighting == probability estimation

Still, it is limited in terms of how it models information needs and queries...

Queries and Information Needs

- A query can represent very different information needs
  – May require different search techniques and ranking algorithms to produce the best rankings

➡ A query can be a poor representation of the information need

  – User may find it difficult to express the information need

  – User is encouraged to enter short queries both by the search engine interface, and by the fact that long queries don't work

Result?

- Interaction with the system occurs
  - during query formulation and reformulation
  - while browsing the result
- Key aspect of effective retrieval
  - users can't change ranking algorithm but can change results through interaction
  - helps refine description of information need
    - » e.g., same initial query, different information needs
    - » how does user describe what they don't know?

ASK Hypothesis

- Belkin et al (1982) proposed a model called Anomalous State of Knowledge
- ASK hypothesis:
  - difficult for people to define exactly what their information need is, because that information is a gap in their knowledge
  - Search engine should look for information that fills those gaps
- Interesting ideas, little practical impact (yet)

Query Expansion

- A variety of *automatic* or *semi-automatic* query expansion techniques have been developed
  - goal is to improve effectiveness by matching related terms
  - semi-automatic techniques require user interaction to select best expansion terms
- Query suggestion is a related technique
  - alternative queries, not necessarily more terms

Relevance Feedback

- User identifies relevant (and maybe non-relevant) documents in the initial result list
- System modifies query using terms from those documents and reranks documents
  - example of ML-based classificiation algorithm to distinguish relevant vs. non-relevant docs
  - but, very little training data
- **Pseudo-relevance feedback** just assumes top-ranked documents are relevant – no user input

Relevance Feedback Example

Top 10 documents for “tropical fish”

1. Badmans **Tropical Fish**

   A freshwater aquarium page covering all aspects of the **tropical fish** hobby. ... to Badman's **Tropical Fish**. ... world of aquariology with Badman's **Tropical Fish**. ...

2. **Tropical Fish**

   Notes on a few species and a gallery of photos of African cichlids.

3. The **Tropical** Tank Homepage - **Tropical Fish** and Aquariums

   Info on **tropical fish** and **tropical** aquariums, large **fish** species index with ... Here you will find lots of information on **Tropical Fish** and Aquariums. ...

4. **Tropical Fish** Centre

   Offers a range of aquarium products, advice on choosing species, feeding, and health care, and a discussion board.

5. **Tropical fish** - Wikipedia, the free encyclopedia

   Tropical fish are popular aquarium **fish** , due to their often bright coloration. ... Practical Fishkeeping • **Tropical Fish** Hobbyist • Koi. Aquarium related companies: ...

6. **Tropical Fish** Find

   Home page for **Tropical Fish** Internet Directory ... stores, forums, clubs, **fish** facts, **tropical fish** compatibility and aquarium ...

7. Breeding **tropical fish**

   ... intrested in keeping and/or breeding **Tropical**, Marine, Pond and Coldwater **fish**. ... Breeding **Tropical Fish** ... breeding **tropical**, marine, coldwater & pond **fish**. ...

8. FishLore

   Includes **tropical** freshwater aquarium how-to guides, FAQs, **fish** profiles, articles, and forums.

9. Cathy's **Tropical Fish** Keeping

   Information on setting up and maintaining a successful freshwater aquarium.

10. **Tropical Fish** Place

    Tropical Fish information for your freshwater **fish** tank ... great amount of information about a great hobby, a freshwater **tropical fish** tank. ...

Relevance Feedback Example

- If we assume top 10 are relevant, most frequent terms are (with frequency):

  a (926), td (535), href (495), http (357), width (345), com (343), nbsp (316), www (260), tr (239), htm (233), class (225), jpg (221)

  » too many stopwords and HTML expressions

- Use only snippets and remove stopwords

  tropical (26), fish (28), aquarium (8), freshwater (5), breeding (4), information (3), species (3), tank (2), Badman’s (2), page (2), hobby (2), forums (2)

Relevance Feedback Example

- If document 7 (“Breeding tropical fish”) is *explicitly* indicated to be relevant, the most frequent terms are:

  breeding (4), fish (4), tropical (4), marine (2), pond (2), coldwater (2), keeping (1), interested (1)

- Specific weights and scoring methods used for relevance feedback depend on retrieval model

Relevance Feedback

- Both relevance feedback and pseudo-relevance feedback are effective, but not used in many applications
  - pseudo-relevance feedback has reliability issues, especially with queries that don’t retrieve many relevant documents
- Some applications use relevance feedback
  - filtering, “more like this”
- Query suggestion more popular
  - may be less accurate, but can work if initial query fails

LMs for Retrieval

- 3 possibilities:

QL – probability of generating the query text from a document language model

– **probability of generating the document text from a query language model**

– comparing the language models representing the query and document topics

- Models of topical relevance

Relevance Models

- ***Relevance model*** – language model representing information need
  - query and relevant documents are samples from this model
- *P(D|R)* - probability of generating the text in a document given a relevance model
  - *document likelihood* model
  - less effective than query likelihood
  - Difficult to calculate and to compare across documents of different lengths

LMs for Retrieval

- 3 possibilities:

QL – probability of generating the query text from a document language model

DL – probability of generating the document text from a query language model

– **comparing the language models representing the query and document topics**

- Models of topical relevance

Pseudo-Relevance Feedback

- Estimate relevance model from query and top-ranked documents
- Rank documents by similarity of document model to relevance model
- ***Kullback-Leibler divergence*** (KL-divergence) is a well-known measure of the difference between two probability distributions

KL-Divergence

- Given the ***true*** probability distribution *P* and another distribution *Q* that is an ***approximation*** to *P*,

$$KL(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

  - Use negative KL-divergence for ranking, and assume relevance model *R* is the true distribution (not symmetric),

$$\sum_{w \in V} P(w|R) \log P(w|D) - \sum_{w \in V} P(w|R) \log P(w|R)$$

KL-Divergence

- Given a simple maximum likelihood estimate for *P(w|R),* based on the frequency in the query text, ranking score is

$$\sum_{w \in V} \frac{f_{w,Q}}{|Q|} \log P(w|D)$$

  - rank-equivalent to query likelihood score
- Query likelihood model is a special case of retrieval based on relevance model

Estimating the Relevance Model

- Probability of pulling a word *w* out of the "bucket" representing the relevance model depends on the *n* query words we have just pulled out

$$P(w|R) \approx P(w|q_1 \ldots q_n)$$

- By definition

$$P(w|R) \approx \frac{P(w, q_1 \ldots q_n)}{P(q_1 \ldots q_n)}$$

Estimating the Relevance Model

- Joint probability is

$$P(w, q_1 \ldots q_n) = \sum_{D \in \mathcal{C}} p(D) P(w, q_1 \ldots q_n|D)$$

- Assume

$$P(w, q_1 \ldots q_n|D) = P(w|D) \prod_{i=1}^{n} P(q_i|D)$$

- Gives

$$P(w, q_1 \ldots q_n) = \sum_{D \in \mathcal{C}} P(D) P(w|D) \prod_{i=1}^{n} P(q_i|D)$$

Estimating the Relevance Model

- *P(D)* usually assumed to be uniform
- *P(w, q_1 . . . q_n)* is simply a weighted average of the language model probabilities for *w* in a set of documents, where the weights are the query likelihood scores for those documents
- Formal model for pseudo-relevance feedback
  - query expansion technique

Ranking based on the Relevance Model

1. Rank documents using the query likelihood score for query Q.
2. Select some number of the top-ranked documents to be the set $\mathcal{C}$.
3. Calculate the relevance model probabilities $P(w|R)$.

4. Rank documents again using the KL-divergence score

$$\sum_w P(w|R) \log P(w|D)$$

Example from Top 10 Docs

president lincoln	*abraham lincoln*	*fishing*	*tropical fish*
lincoln	lincoln	fish	fish
president	america	farm	tropic
room	president	salmon	japan
bedroom	faith	new	aquarium
house	guest	wild	water
white	abraham	water	species
america	new	caught	aquatic
guest	room	catch	fair
serve	christian	tag	china
bed	history	time	coral
washington	public	eat	source
old	bedroom	raise	tank
office	war	city	reef
war	politics	people	animal
long	old	fishermen	tarpon
abraham	national	boat	fishery

Example from Top 50 Docs

president lincoln	*abraham lincoln*	*fishing*	*tropical fish*
lincoln	lincoln	fish	fish
president	president	water	tropic
america	america	catch	water
new	abraham	reef	storm
national	war	fishermen	species
great	man	river	boat
white	civil	new	sea
war	new	year	river
washington	history	time	country
clinton	two	bass	tuna
house	room	boat	world
history	booth	world	million
time	time	farm	state
center	politics	angle	time
kennedy	public	fly	japan
room	guest	trout	mile