DAT630
Queries and Information Needs

Search Engines, Chapter 6

19/10/2016

Krisztian Balog | University of Stavanger

Information Needs

- An *information need* is the underlying cause of the query that a person submits to a search engine
  - Sometimes called *query intent*
- Categorized using variety of dimensions
  - E.g., number of relevant documents
  - Type of information that is needed
  - Type of task that led to the requirement for information

Queries

- *Keyword queries*: simple, natural language queries, designed to enable everyone to search
- Typical query length in web search is 2.3 words
- Keyword selection is not always easy
  - Query refinement techniques can help

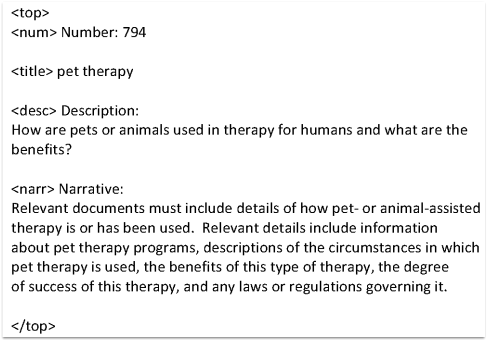
Query vs. Information Need

"I would like to have a test drive before I buy the Kawasaki ER6f"
Information need
Query
Google
kawasaki stavanger

Query vs. Information Need

- 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 often don't work very well

TREC Topic Example



```
<top>
<num> Number: 794

<title> pet therapy

<desc> Description:
How are pets or animals used in therapy for humans and what are the
benefits?

<narr> Narrative:
Relevant documents must include details of how pet- or animal-assisted
therapy is or has been used.  Relevant details include information
about pet therapy programs, descriptions of the circumstances in which
pet therapy is used, the benefits of this type of therapy, the degree
of success of this therapy, and any laws or regulations governing it.

</top>
```

Query Reformulation

- Rewrite or transform original query to better match underlying intent
- Can happen implicitly or explicitly (suggestion)
- Many techniques, including
  - Spelling correction
  - Query expansion
  - Query suggestion
  - Relevance feedback

Spelling Correction

- Important part of query processing
  - 10-15% of all web queries have spelling errors
- Errors include typical word processing errors but also many other types, e.g.,

poiner sisters
brimingham news
catamarn sailing
hair extenssions
marshmellow world
miniture golf courses
psyhics
home doceration

realstateisting.bc.com
akia 1080i manunal
ultimatwarcade
mainscourcebank
dellottitouche

Spelling Correction

- Basic approach: suggest corrections for words that are not in a spelling dictionary
- Suggestions found by comparing word to dictionary words using similarity measure
- Most common similarity measure is edit distance
  - Number of operations required to transform one word into the other

Edit Distance

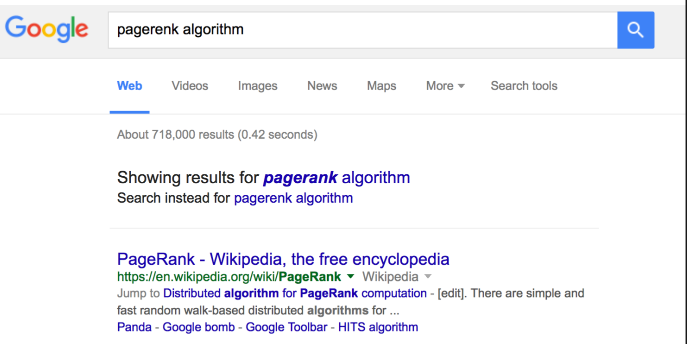
- Damerau-Levenshtein distance
  - Counts the minimum number of insertions, deletions, substitutions, or transpositions of single characters required
  - E.g., Damerau-Levenshtein distance 1

extenssions → extensions(insertion error)
poiner → pointer (deletion error)
marshmellow → marshmallow (substitution error)
brimingham → birmingham (transposition error)

  - distance 2

doceration → deceration
deceration → decoration

Spelling Correction



About 718,000 results (0.42 seconds)

Showing results for ***pagerank*** algorithm
Search instead for pagerenk algorithm

PageRank - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**PageRank** ▾ Wikipedia ▾
Jump to Distributed **algorithm** for **PageRank** computation - [edit]. There are simple and fast random walk-based distributed **algorithms** for ...
Panda - Google bomb - Google Toolbar - HITS algorithm

Query Expansion

- Early search engines used thesauri
  - Adding synonyms or more specific terms using query operators based on a thesaurus
  - Improves search effectiveness (if used correctly)
- Modern approaches are usually based on an analysis of term co-occurrence
  - Either in the entire document collection, a large collection of queries, or the top-ranked documents in a result list

Term Association Measures

- Various statistical measures to estimate the strength of the association between two terms

Measure	*Formula*
Mutual information (MIM)	$\frac{n_{ab}}{n_a.n_b}$
Expected Mutual Information ($EMIM$)	$n_{ab}.\log(N.\frac{n_{ab}}{n_a.n_b})$
Chi-square (χ^2)	$\frac{(n_{ab}-\frac{1}{N}.n_a.n_b)^2}{n_a.n_b}$
Dice's coefficient ($Dice$)	$\frac{n_{ab}}{n_a+n_b}$

Term Association Examples

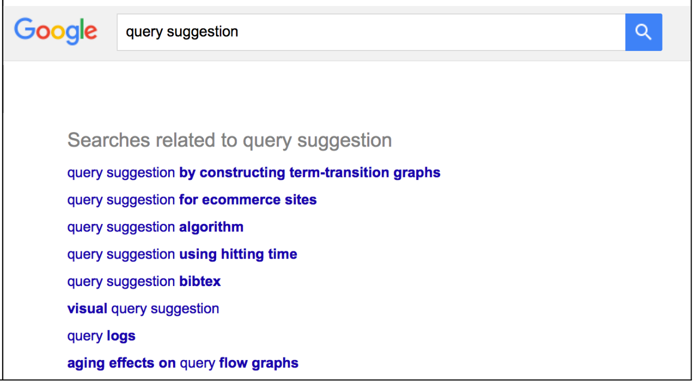
MIM	*EMIM*	χ^2	*Dice*
trmm	forest	trmm	forest
itto	tree	itto	exotic
ortuno	rain	ortuno	timber
kuroshio	island	kuroshio	rain
ivirgarzama	like	ivirgarzama	banana
biofunction	fish	biofunction	deforestation
kapiolani	most	kapiolani	plantation
bstilla	water	bstilla	coconut
almagreb	fruit	almagreb	jungle
jackfruit	area	jackfruit	tree
adeo	world	adeo	rainforest
xishuangbanna	america	xishuangbanna	palm
frangipani	some	frangipani	hardwood
yuca	live	yuca	greenhouse
anthurium	plant	anthurium	logging

Most strongly associated words for "tropical" in a collection of TREC news stories. Co-occurrence counts are measured at the document level.

Query Suggestion

- Explicit query reformulation by the user
- The search engine suggests alternative queries (not necessarily more terms) based on search query logs

Query Suggestion



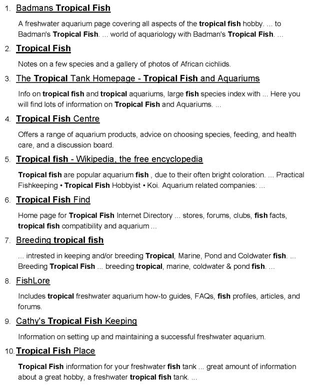
Searches related to query suggestion

query suggestion **by constructing term-transition graphs**
query suggestion **for ecommerce sites**
query suggestion **algorithm**
query suggestion **using hitting time**
query suggestion **bibtex**
visual query suggestion
query **logs**
aging effects on query **flow graphs**

Relevance Feedback

- User identifies relevant (and maybe non-relevant) documents in the initial result list
- System modifies the query using terms from those documents and re-ranks documents
- *Pseudo-relevance feedback* just assumes top-ranked documents are relevant – no user input is required

Relevance Feedback Example



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. ...

Top 10 documents for "tropical fish"

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
  - E.g., "more like this"
- Query suggestion is more popular

Query Models in LM scoring

- Standard log- query likelihood scoring

$$\log P(d|q) \propto \underbrace{\log P(q|d)}_{\downarrow} + \log P(d)$$

Frequency of the term in the query

$$logP(q|d) = \sum_{t\in q} \boxed{f_{t,q}} \cdot \log P(t|\theta_d)$$

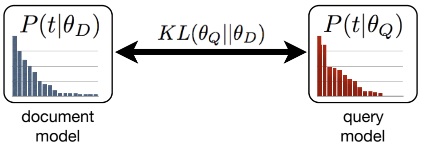
$$\Big\downarrow \textit{replace}$$

$$logP(q|d) = \sum_{t\in q} \boxed{P(t|\theta_q)} \cdot \log P(t|\theta_d)$$

Represent the query as a distribution over terms (i.e., query LM)

Alternatively

- Assuming uniform document priors, it provides the same ranking as minimizing the KL-divergence between two probability distributions



Relevance Models

[Lavrenko and Croft, 2001]

- Using the joint probability of observing t with query terms in feedback documents
  - Feedback documents may be obtained using either explicit or pseudo relevance feedback

$$p(t|\hat{q}) \approx \frac{p(t, q_1, \ldots, q_n)}{\sum_{t'} p(t', q_1, \ldots, q_n)}$$

- **RM1**(all query terms are conditioned on t)

$$p(t, q_1 \ldots q_k) = \sum_{d \in M} p(d) \cdot p(t|d) \prod_{i=1}^{k} p(q_i|d)$$

- **RM2** (pairwise independence assumption)

$$p(t, q_1 \ldots q_k) = p(t) \prod_{i=1}^{k} \sum_{d \in M} p(d|t) \cdot p(q_i|d)$$