CS4300 (2013FA)
Query Refinement

Jon Park
DEPARTMENT OF COMPUTER SCIENCE
CORNELL UNIVERSITY

Content

- Why Refine Queries?
- Query Reduction
  - Stopword Removal
- Query Expansion
  - Thesaurus
  - Query-Based Stemming
    - Stem Classes
    - Association Measures
- ~~Etc.~~
  - ~~Spell Checking, Personalization, Rel. Feedback~~

1. Why Refine Queries?

1. Why Refine Queries?

1. Why Refine Queries?

- Natural Language Query
  - Hmm.. I wonder where CS4300 is being held..
  - Where is CS4300 being held?
- Keyword Query
  - CS4300 where
  - cornell CS4300 information retrieval location

2. Query Reduction

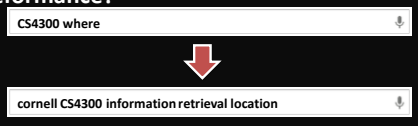
2. Query Reduction

- Queries may contain words that do not help identifying relevant documents
  - Common words
  - Purely functional words
- Stopword Removal
  - While Indexing?
    - "to be or not to be" or "Just a Taste"
  - While Querying?

3. Query Expansion

3. Query Expansion

- Can we expand queries to improve the peformance?



- The key is to add words appropriate for the topic to the query
  - e.g. "aquarium" as an expansion term for "tank" in "tropical fish tanks"? "armor for tanks"?

3.1. Thesaurus

- Used in early search engines as a tool for manual indexing (tagging) and query formulation
  - specified preferred terms and relationships between them
  - also called *controlled vocabulary*
- Currently, automatic query expansion using general purpose thesaurus have not been effective.
  - Synonyms w.r.t. many different meanings

3.2. Query-Based Stemming

- Stemming can be useful for reducing variations in words, but is imperfect.
  - {"fish", "fished", "fishing"} → "fish"
  - {"bank", "banked", "banking", "bankings", "banks"} → "bank"
- Query-Based Stemming
  - Stemming is done only on the query side
    - How about Indexing side?
  - Query is expanded with word variants
    - e.g. "rock climbing" → "rock climbing climb"

3.2.1. Stem Classes

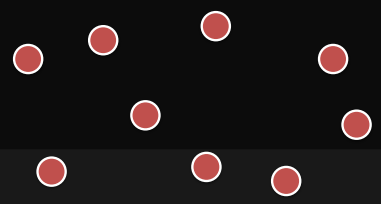
- A *stem class* is the group of words that will be transformed into the same stem by the stemming algorithm
  - bank, banked, banking, bankings, banks
  - ocean, oceaneering, oceanic, oceanics, oceanizatino, oceans
  - polic, polical, polically, police, policeable, policed, policement, policer, policers, polices, policial, policially, policier, policiers, policies, policing, policization, policize, policly, policy, policying, policys

3.2.1. Stem Classes

- Issues with Stem Classes
  - Inaccurate :Inflection vs derivation
- Solution
  - Split stem classes into smaller sets using word co-occurrence information (association measure)
    - Algorithm: Given a stem class,

      1. Build an edgeless graph whose vertices are words in the given stem class
      2. Connect 2 words with an edge iff the association score is above a threshold.
      3. Set each connected component as its own cluster

3.2.1. Stem Classes



Term Association Measures

- *Dice's Coefficient*

$$\frac{2.n_{ab}}{n_a+n_b} \overset{rank}{=} \frac{n_{ab}}{n_a+n_b}$$

- *Mutual Information*

$$\log \frac{P(a,b)}{P(a)P(b)} = \log N.\frac{n_{ab}}{n_a.n_b} \overset{rank}{=} \frac{n_{ab}}{n_a.n_b}$$

  - *N* number of text windows in the collection
  - *P(a)* probability that word *a* occurs in a given window of text
  - *P(a, b)* probability that *a* and *b* occur in the same window of text
  - Measures the extent to which 2 words occur independently

Term Association Measures

- Mutual Information measure favors low frequency terms
- *Expected Mutual Information* (EMI)

$$P(a,b).\log\frac{P(a,b)}{P(a)P(b)} = \frac{n_{ab}}{N}\log(N.\frac{n_{ab}}{n_a.n_b}) \overset{rank}{=} n_{ab}.\log(N.\frac{n_{ab}}{n_a.n_b})$$

  - actually only 1 part of full EMI, focused on word occurrence

Term Association Measures

- *Pearson's Chi-squared (χ^2) measure*
  - compares the number of co-occurrences of two words with the expected number of co-occurrences if the two words were independent
  - normalizes this comparison by the expected number
  - also limited form focused on word co-occurrence

$$\frac{(n_{ab}-N.\frac{n_a}{N}.\frac{n_b}{N})^2}{N.\frac{n_a}{N}.\frac{n_b}{N}} \overset{rank}{=} \frac{(n_{ab}-\frac{1}{N}.n_a.n_b)^2}{n_a.n_b}$$

3.3. Association Measures

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 ($Dice$)	$\frac{n_{ab}}{n_a+n_b}$