Query Refinement

Jon Park
DEPARTMENT OF COMPUTER SCIENCE
CORNELL UNIVERSITY
CS4300 (2013FA)

1. Why Refine Queries?

1. Why Refine Queries?

Hmm.. I wonder where CS4300 is being held..

Where is CS4300 being held?

CS4300 where

cornell CS4300 information retrieval location

2. Query Reduction

• Natural Language Query

Hmm.. I wonder where CS4300 is being held..

Where is CS4300 being held?

CS4300 where

cornell CS4300 information retrieval location

• Keyword Query

CS4300 where

cornell CS4300 information retrieval location

Content

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

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

- Can we expand queries to improve the performance?
  - CS4300 where
  - Cornell CS4300 Information retrieval location
- 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”) \(\rightarrow\) “fish”
  - (“bank”, “banked”, “banking”, “bankings”, “banks”) \(\rightarrow\) “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” \(\rightarrow\) “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, oceanneering, oceanic, oceanics, oceanization, oceans
  - polic, polical, polically, police, policeable, policed, policeman, policer, policers, polices, policial, policially, policiers, policiers, policies, policing, policization, policize, policy, policying, policies
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

\[ \frac{2n_{ab}}{n_a + n_b} \]

• Mutual Information

\[ \log \frac{P(a,b)}{P(a)P(b)} = \log \frac{n_{ab}}{n_a n_b} \]

– \( N \) 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

• Pearson’s Chi-squared (\( \chi^2 \)) measure

\[ \frac{(n_{ab} - \frac{n_a n_b}{N})^2}{\frac{n_a}{N} \cdot \frac{n_b}{N}} \]

– 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

3.3. Association Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutual Information (MIM)</td>
<td>( \frac{n_{ab} - \frac{n_a n_b}{N}}{\sqrt{n_a n_b \cdot \frac{N}{n_a + n_b}}} )</td>
</tr>
<tr>
<td>Expected Mutual Information (EMIM)</td>
<td>( \frac{n_{ab} \cdot \log(N \cdot \frac{n_{ab}}{n_a + n_b})}{n_a + n_b} )</td>
</tr>
<tr>
<td>Chi-square (( \chi^2 ))</td>
<td>( \frac{(n_{ab} - \frac{n_a n_b}{N})^2}{\frac{n_a}{N} \cdot \frac{n_b}{N}} )</td>
</tr>
<tr>
<td>Dice’s coefficient (Dice)</td>
<td>( \frac{n_{ab}}{n_a + n_b} )</td>
</tr>
</tbody>
</table>