Idea: Query Expansion

- Users often start with short queries with ambiguous representations
- Observation: Many people refine their queries by analyzing the results from initial queries, or consulting other resources (thesaurus)
  - By adding and removing terms
  - By reweighting terms
  - By adding other features (e.g., Boolean operators)
- Technique of query expansion:
  Can a better query be created automatically?
Idea: Query Expansion

• Add terms to query to improve recall
  – And possibly precision
• Initial intuition: Help users find synonyms for query terms
  – Later: Help users find good query terms
• Query Expansion via External Resources
  – Thesaurus
    • “Industrial Chemical Thesaurus”, “Medical Subject Headings” (MeSH)
  – Semantic network
    • WordNet
Query Expansion via External Resources: Thesaurus

Word: Bank (Institution)
coffer, countinghouse, credit union, depository, exchequer, fund, hoard, investment firm, repository, reserve, reservoir, safe, savings, stock, stockpile...

Word: Bank (Ground)
beach, berry bank, caisse populaire, cay, cliff, coast, edge, embankment, lakefront, lakeshore, lakeside, ledge, levee, oceanfront, reef, riverfront, riverside, ...

Word: Java (Coffee)
Jamocha, cafe, cafe noir, cappuccino, decaf, demitasse, dishwater, espresso...

Word: Refusal
abnegation, ban, choice, cold shoulder*, declension, declination, defiance, disallowance, disapproval, disavowal, disclaimer,

---

Query Expansion via External Resources: Thesaurus

<table>
<thead>
<tr>
<th>MeSH Heading</th>
<th>Neoplasms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree Number</td>
<td>C04</td>
</tr>
</tbody>
</table>

Annotation:
avoid: too general, prefer specifics; policy: Manual section 24; / chem ind permitted but consider also CARCINOGENS; / class: consider also NEOPLASM STAGING (see note there) but "grading" = / pathol; / chol: consider also ONCOGENIC VIRUSES; / vet: Manual 24.6+ or TN 136....

Scope Note:
New abnormal growth of tissue. Malignant neoplasms show a greater degree of anaplasia and have the properties of invasion and metastasis, compared to benign neoplasms.

Entry Term: Cancer
Entry Term: Tumors
Entry Term: Benign Neoplasm
Entry Term: Neoplasms, Benign
Query Expansion via External Resources: Semantic Network

- WordNet: a lexical thesaurus organized into 4 taxonomies by part of speech (George Millet et al.)
- Inspired by psycholinguistic theories of human lexical memory
- English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one concept
- Multiple relations link the synonym sets
  - Hyponyms: \( Y \) is a hyponym of \( X \) if every \( Y \) is a (kind of) \( X \)
  - Hypernyms: \( Y \) is a hypernym of \( X \) if every \( X \) is a (kind of) \( Y \)
  - Meronyms: \( Y \) is a meronym of \( X \) if \( Y \) is a part of \( X \)
  - Holonyms: \( Y \) is a holonym of \( X \) if \( X \) is a part of \( Y \)
Query Expansion via External Resources: Semantic Network

• Three sense of the noun “Java”
  1. Java (an island in Indonesia south of Borneo; one of the world's most densely populated regions)
  2. java (a beverage consisting of an infusion of ground coffee beans) "he ordered a cup of java"
  3. Java (a simple platform-independent object-oriented programming language used for writing applets that are downloaded from the World Wide Web by a client and run on the client's machine)

• The hypernym of Sense 3 of “Java”
  =>: (n) object-oriented programming language, object-oriented programming language
  =>: (n) programming language, programming language
  =>: (n) artificial language
  =>: (n) language, linguistic communication
  =>: (n) communication
  =>: (n) abstraction
  =>: (n) abstract entity
  =>: (n) entity
Query Expansion via External Resources: Semantic Network

• The meronym of Sense 1 of “Java”
  • =>: (n) Jakarta, Djakarta, capital of Indonesia (capital and largest city of Indonesia; located on the island of Java; founded by the Dutch in 17th century)
  • =>: (n) Bandung (a city in Indonesia; located on western Java (southeast of Jakarta); a resort known for its climate)
  • =>: (n) Semarang, Samarang (a port city is southern Indonesia; located in northern Java)

Query Expansion via External Resources: Semantic Network

• User select synonym sets for some query terms
  – Add to query all synonyms in synset
  – Add to query all hypernyms (“… is a kind of X”) up to depth n
  – May add hyponyms, meronym etc
• Query expansions with WordNet has not been consistently useful
  – What to expand? To what kind of detail?
  – Not query-specific, difficult to disambiguate the senses
  – some positive results reported using conservative set of synonyms close to limited query terms