CS47300: Web Information Search and Management

Query Expansion
Prof. Chris Clifton
25 September 2020
Material adapted from course created by Dr. Luo Si, now leading Alibaba research group

Retrieval Models

? Information Need

Representation

Query

Retrieval Model

Indexed Objects

Retrieved Objects

Evaluation/Feedback
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?
Query Expansion

Query Expansion
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>
<tr>
<td>Annotation</td>
<td>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; / etiol; consider also ONCOGENIC VIRUSES; / vet- Manual 24.6+ or TN 136....</td>
</tr>
<tr>
<td>Scope Note</td>
<td>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.</td>
</tr>
<tr>
<td>Entry Term</td>
<td>Cancer</td>
</tr>
<tr>
<td>Entry Term</td>
<td>Tumors</td>
</tr>
<tr>
<td>Entry Term</td>
<td>Benign Neoplasms</td>
</tr>
<tr>
<td>Entry Term</td>
<td>Neoplasms, Benign</td>
</tr>
</tbody>
</table>

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 senses 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)
Query Expansion via External Resources: Semantic Network

• 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

Idea: Query Expansion

- Add terms to query to improve recall
  - And possibly precision
- Query Expansion via External Resources
  - Thesaurus
    - “Industrial Chemical Thesaurus”, “Medical Subject Headings” (MeSH)
  - Semantic network
    - WordNet
- Relevance Feedback
  - Use user-specified “good documents” to get new terms
  - Blind/Pseudo Relevance Feedback
Query Expansion via Relevance Feedback

Retrieval Models

- Representation
- Query
- Retrieval Model
- Retrieved Objects
- Evaluation/Feedback
- Indexed Objects

Information Need
Query Expansion: Relevance Feedback

Query: iran iraq war

Initial Retrieval Result

1. 0.643 07/11/88, Japan Aid to Buy Gear For Ships in Persian Gulf
2. 0.582 08/21/90, Iraq's Not-So-Tough Army
3. 0.569 09/10/90, Societe Generale Iran Pact
4. 0.566 08/11/88, South Korea Estimates Iran-Iraq Building Orders
5. 0.562 01/02/92, International: Iran Seeks Aid for War Damage
6. 0.541 12/09/86, Army Suspends Firings Of TOWs Due to Problems

New query representation:

10.82 Iran 9.54 iraq 6.53 war
2.3 army 3.3 perisan 1.2 aid
1.5 gulf 1.8 raegan 1.02 ship
1.61 troop 1.2 military 1.1 damage
Query Expansion: Relevance Feedback

Updated Query

Refined Retrieval Result

+1  0.547  08/21/90, Iraq's Not-So-Tough Army
+2  0.529  01/02/92, International: Iran Seeks Aid for War Damage
  3  0.515  07/11/88, Japan Aid to Buy Gear For Ships in Persian Gulf
  4.  0.511  09/10/90, Societe Generale Iran Pact
  5  0.509  08/11/88, South Korea Estimates Iran-Iraq Building Orders
+ 6.  0.498  06/05/87, Reagan to Urge Allies at Venice Summit To Endorse Cease-Fire in Iran-Iraq War

Relevance Feedback Vector Space Model

- Two types of words are likely to be included in the expanded query
  - Topic specific words: good representative words
  - General words: introduce ambiguity into the query, may lead to degradation of the retrieval performance
  - Utilize both positive and negative documents to distinguish representative words
Relevance Feedback Vector Space Model

- **Desirable weights for $\alpha$ and $\beta$**

  Try find $\alpha$ and $\beta$ such that
  
  $\tilde{q}(\alpha, \beta) \cdot \vec{d}_i \geq 1$ for $\vec{d}_i \in R$
  $\tilde{q}(\alpha, \beta) \cdot \vec{d}_i \leq -1$ for $\vec{d}_i \in NR$

Relevance Feedback Vector Space Model

- **Goal**: Move new query close to relevant documents and far away from irrelevant documents

- **Approach**: New query is a weighted average of original query, and relevant and non-relevant document vectors

  \[
  \tilde{q}' = \tilde{q} + \alpha \frac{1}{|R|} \sum_{\vec{d}_i \in R} \vec{d}_i - \beta \frac{1}{|NR|} \sum_{\vec{d}_i \in NR} \vec{d}_i \quad \text{(Rocchio formula)}
  \]
Relevance Feedback Vector Space Model

- **Goal**: Move new query close to relevant documents and far away from irrelevant documents
- **Approach**: New query is a weighted average of original query, and relevant and non-relevant document vectors

\[
\vec{q}' = \vec{q} + \alpha \frac{1}{|R|} \sum_{d_i \in R} \vec{d}_i - \beta \frac{1}{|NR|} \sum_{d_i \in NR} \vec{d}_i \quad \text{(Rocchio formula)}
\]

How do we set the desired weights?

Relevance Feedback Vector Space Model

- Desirable weights for \(\alpha\) and \(\beta\)
- Exhaustive search
- Heuristic choice
  \(\alpha=0.5; \quad \beta=0.25\)
- Learning method
  - Perceptron algorithm (Rocchio)
  - Support Vector Machine (SVM)
  - Regression
  - Neural network algorithm
Blind (Pseudo) Relevance Feedback

- What if users only mark some relevant documents?
  - Use bottom documents as negative documents
- What if users only mark some irrelevant documents?
  - Use top documents in initial ranked lists and queries as positive documents
- What if users do not provide any relevance judgments?
  - Use top documents in initial ranked lists as positive documents; bottom documents as negative documents
- What about implicit feedback?
  - Use reading time, scrolling and other interaction?

Blind (Pseudo) Relevance Feedback

Approaches
- Pseudo-relevance feedback
  - Assume top N (e.g., 20) documents in initial list are relevant
  - Assume bottom N’ (e.g., 200-300) in initial list are irrelevant
  - Calculate weights of term according to some criterion (e.g., Rocchio)
  - Select top M (e.g., 10) terms
- Local context analysis
  - Similar approach to pseudo-relevance feedback
  - But use passages instead of documents for initial retrieval; use different term weight selection algorithms
Relevance Feedback Summary

• Relevance feedback can be very effective
• Effectiveness depends on the number of judged documents (positive documents more important)
• An area of active research (many open questions)
• Effectiveness also depends on the quality of initial retrieval results (what about bad initial results?)
• Need to do retrieval process twice

Summary: Query Expansion

• Add terms to query to improve recall
  – And possibly precision
• Query Expansion via External Resources
  – Thesaurus
    • “Industrial Chemical Thesaurus”, “Medical Subject Headings” (MeSH)
  – Semantic network
    • WordNet
• Relevance Feedback
  – Use user-specified “good documents” to get new terms
  – Blind/Pseudo Relevance Feedback
    • Rocchio