CS47300: Web Information Search and Management

Relevance Feedback
Prof. Chris Clifton
15 September 2017
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
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

- **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

$$
q' = \bar{q} + \alpha \frac{1}{|R|} \sum_{d_i \in R} d_i - \beta \frac{1}{|NR|} \sum_{d_i \in NR} d_i \quad \text{(Rocchio formula)}
$$

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

Try find \( \alpha \) and \( \beta \)
such that

\[
\tilde{q}(\alpha, \beta) \cdot \dd_i \geq 1 \text{ for } \dd_i \in R \\
\tilde{q}(\alpha, \beta) \cdot \dd_i \leq -1 \text{ for } \dd_i \in NR
\]
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