CS317

Information Retrieval

Week 08

Muhammad Rafi

April 01, 2021

Relevance Feedback & Query Expansion

Chapter No. 9

Agenda

- Exercises

Problem 9.1

- In Rocchio's algorithm, what weight setting for α/β/γ does a "Find pages like this one" search correspond to?
  - "Find pages like this one" ignores the query and no negative judgments are used.
  - A good positive judgement is available (Example Page)
  - Hence alpha = gamma = 0. This implies beta = 1.

Problem 9.2

- Give three reasons why relevance feedback has been little used in web search?

Relevance Feedback is mainly used to increase recall, but web users are mainly concerned about the precision of the top few results.
Relevance Feedback slows down returning results as you need to run two sequential queries, the second of which is slower to compute than the first. Web users hate to be kept waiting.
Relevance Feedback is one way of dealing with alternate ways to express an idea (synonymy), but indexing anchor text is commonly already a good way to solve this problem.
Relevance Feedback complicates the user interface.
Relevance Feedback is difficult to explain to a common user.

Problem 9.3

- Give three reasons why relevance feedback has been little used in web search?
  - RF slows down returning results as you need to run two sequential queries, the second of which is slower to compute than the first. Web users hate to be kept waiting.
  - RF is mainly used to increase recall, but web users are mainly concerned about the precision of the top few results.
  - RF is one way of dealing with alternate ways to express an idea (synonymy), but indexing anchor text is commonly already a good way to solve this problem.
  - RF complicates the user interface.
  - RF is difficult to explain.

Problem 9.b

- What are some of the problems associated with relevance feedback?
  - Relevance feedback is expensive
  - Relevance feedback creates long modified queries
  - Long queries are expensive to process
  - Users are reluctant to provide explicit feedback
  - Its often hard to understand why a particular document was retrieved after applying relevance feedback

Problem 9.c

- Discuss the pros and cons of implicit (indirect) vs. explicit (direct) feedbacks ?

Implicit(Indirect) feedback	Explicit (Direct) feedback
Implicit (indirect) feedback does not bother user for explicit actions. It is fast and can be possible for large IR system. It is less reliable and possibly introduce a problem of query drift.	Explicit (Direct) feedback requires user to marks document relevant. It is slow process and does not scale to large systems. It is more reliable and generally save from query drift.

Problem 9.d

- What is meant by query expansion? Give an example query that need expansion? When query expansion is very useful?
  - Query expansion is an autonomous process of reformulating a seed query (q_o) to improve retrieval performance in information retrieval systems.
  - It is generally performed to bridge the gap between user information need and the posed query.
  - A general case is a lexical mismatch for example astronauts or cosmonauts are mean the same thing, the query for astronauts implicitly union with the term cosmonauts to bridge the gap and get the relevant documents from both the terms.
  - It is very useful technique for such a situation.

Problem

Suppose that a user's initial query is q= w1 w3 w2 and IR systems return four documents. User selected d1= w2 w3 w4, d2= w3 w3 w4 w5 and d4= w1 w3 w4 w1 as relevant. While d3= w2 w4 w5 w3 as non-relevant to her query. Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors. Using Rocchio relevance feedback algorithm to get modify query vector (optimal) after relevance feedback? Rocchio equation is given below.

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j\in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j\in D_{nr}}\vec{d}_j$$

$q = < 1, 1, 1, 0, 0>$
$d_1 = < 0, 1, 1, 1, 0>$
$d_2 = < 0, 0, 2, 1, 1>$
$d_3 = < 0, 1, 1, 1, 1>$
$d_4 = < 2, 0, 1, 1, 0>$

Let's α=0.1, β=0.6, and γ=0.3

Problem

Suppose that a user's initial query is q= w1 w3 w2 and IR systems return four documents. User selected d1= w2 w3 w4, d2= w3 w3 w4 w5 and d4= w1 w3 w4 w1 as relevant. While d3= w2 w4 w5 w3 as non-relevant to her query. Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors. Using Rocchio relevance feedback algorithm to get modify query vector (optimal) after relevance feedback? Rocchio equation is given below.

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$$

Using the given equation, we will get,

q_m= α * < 1, 1, 1, 0, 0> + β* 1/3 *{ < 0, 1, 1, 1, 0> + < 0, 0, 2, 1, 1> + <2,0,1,1,0>} – γ *{< 0, 1, 1, 1, 1> }

q_m= (0.1) * < 1, 1, 1, 0, 0> +(0.6)* 1/3 *{ < 0, 1, 1, 1, 0> + < 0, 0, 2, 1, 1> + <2,0,1,1,0>} – (0.1) *{< 0, 1, 1, 1, 1> }