National University of Computer & Emerging Sciences

Midterm Examination II – Fall 2015 -sol

Course: IR&TM (CS567) Time Allowed: 1 Hour

Date: October 19, 2015 Max. Marks: 40

Instructions: Attempt all question. Be to the point. Draw neat and clean diagram/code where necessary. Answer each question on the new page of the answer book, no marks for junk explanations. You must address all inquires in a question.

Question No. 1 **[Time: 20 Min] [Marks: 14]**

Answer the following questions briefly using 4-5 lines of answer book. Be precise, accurate and to the point, only answer genuine query in the question. Each question is of 2 marks.

a. What is a relevance feedback? Explain the general procedure of relevance feedback in information retrieval.

The idea of relevance feedback (RF) is to involve the user in the retrieval process so as to improve the final result set. In particular, the user gives feedback on the relevance of documents in an initial set of results. The basic procedure is:

- The user issues a (short, simple) query.
- The system returns an initial set of retrieval results.
- The user marks some returned documents as relevant or non-relevant.
- The system computes a better representation of the information need based on the user feedback.
- The system displays a revised set of retrieval results.

b. When does relevance feedback works? Give example situations.

Relevance feedback works only with the following assumptions:

* Users has sufficient knowledge about what they are looking for, which enable them to put an initial query.
* The relevance feedback works in the situation when the relevant documents are similar each other

c. Why is positive feedback likely to be more useful than negative feedback to an IR system?

The idea of feedback system to tell the IR system which documents are relevant to the user and to maximize the return of such documents. Only the positive feedback (annotating the relevant documents) may help to optimize such a situation. If we also provide negative feedback(annotating non-relevant documents) this might pose an orthogonal optimization and thus reduce the precision of the system. Hence system only support positive feedback.

d. What are the advantages of Probabilistic Model of information retrieval over Vector Space Model?

In the Boolean or vector space models of IR, matching is done in a formally defined but semantically imprecise calculus of index terms. Given only a query, an IR system has an uncertain understanding of the information need. Given the query and document representations, a system has an uncertain guess of whether a document has content relevant to the information need or not. Probability theory provides a principled foundation for such reasoning under uncertainty. The Probabilistic model of information retrieval estimate a probabilistic weight of each term with respect to a query, hence it model the impreciseness with probability estimates and it exploit this foundation to estimate how likely it is that a document is relevant to an information need.

e. What are the main assumptions of Probability Ranking Principle (PRP)?

A1: One Random variable for each term(word).
A2: d(w) are mutually independent given R.
A3: P(0/R=1)=P(0/R=0)=0
A4: If the word is not in the query, it is equally likely to occur in relevant and non-relevant populations(documents). practically: We only need to calculate probabilities of common words in query and documents.
A5: On average, a query word will occur in half the relevant documents.
Practical: pw and (1-pw) will cancel out.
A6: Non-relevant set approximated by collection as a whole. (Most documents are non-relevant).

f. How Language Model of Information Retrieval is different from Probabilistic Model?

The language model of IR assume that every document is generated by a distinct model hence its terms probabilities are calculated as per this assumption. While probabilistic model use probabilities for the entire corpus. The language model is more impervious towards the query and better rank the documents for a given query. It is the probability that the query is generated keeping in mind the model of the document.

g. Why a bi-gram language model is practically more challenging? Explain.

A bi-gram language model used bi-gram phrases as a feature for information retrieval. let's assume that a document and query is parsed as a bi-gram feature we need to calculate the probabilities under a language model. for example P(t1,t2,t3) would be P(t1)*P(t2/t1)*P(t3/t2) finding these probabilities from the collection is quite challenging and computational expensive. Hence these models are not practical yet.

Question No. 2	**[Time: 10 Min] [Marks: 6]**

a. Suppose that a user's initial query is "cheap CDs cheap DVDs extremely cheap CDs". The user examines two documents, d1 and d2. She judges d1, with the content "CDs cheap software cheap CDs" relevant and d2 with content "cheap thrills DVDs" non relevant. 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, what would the revised query vector be after relevance feedback? Assume $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$. [4]

First construct the table for the given documents and query:

Terms	**d1**	**d2**	**q**
CDs	2	0	2
cheap	2	1	3
DVDs	0	1	1
extremely	0	0	1
software	1	0	0
thrills	0	1	0

we have for Rocchio's equation for modified query as

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

using the given values we will get

q_m = 1 <2,3,1,1,0,0> + 0.75 <2,2,0,0,1,0> - 0.25<0,1,1,0,0,1>

simplifying, we will get

q_m = <3.5,4.25,0.75,1,0.75,-.25> we will set the negative weight of vector term to zero hence

q_m = <3.5,4.25,0.75,1,0.75,0>

b. In Rocchio's algorithm, what weight setting for $\alpha/\beta/\gamma$ does a "Find pages like this one" search correspond to? [2]

In order to find "Find pages like this one", we need to set β to a very high value as it correspond to relevancy of the search document. We can set α and γ to a minimal values. one such weight assignment would be $\alpha=0$; $\beta=1$; and $\gamma=0$;

Question No. 3	**[Time: 25 Min] [Marks: 20]**

a. Consider a document collection D={d1,d2,d3,d4}, where each document is given below:

d1: dil dil pakistan, jan jan pakistan - (R)
d2: pakistan hum sub ki jan - (NR)
d3: dil aur jan pakistan pakistan - (R)
d4: dil pakistan, jan pakistan - (R)

Assuming that an information system uses Probability Ranking Principle (PRP), with prior probabilities P(R)=0.75 and P(NR)=0.25 to rank the given documents for a fixed query given below:

q: dil jan pakistan

Show all intermediate steps, calculation, tables for PRP and also provide the decreasing rank order of the documents in relevance. [15]

First construct the table for the terms along with the probabilities p(w) and q(w).

	aur	**dil**	**hum**	**jan**	**pakistan**	**sub**	**ki**
$N_1(w)$	1	3	0	3	3	0	0
$N_0(w)$	0	0	1	1	1	1	1
p(w)	1.5/4	3.5/4	0.5/4	3.5/4	3.5/4	0.5/4	0.5/4
q(w)	0.5/2	0.5/2	1.5/2	1.5/2	1.5/2	1.5/2	1.5/2

we know that a RVS for a document d_i we have

RVS $d_i = P(R=1/ d_i, q) = P(R) * \prod_{\text{(over common terms)}} (p_w(1-q_w)) / (q_w(1-p_w))$

so
RVS $d_1 = 85.75$
RVS $d_2 = 4.0833$
RVS $d_3 = 85.75$
RVS $d_4 = 85.75$

so, the rank order is
d_1, d_3, d_4
d_2

b. If a new document **d5** is given, How can we decide whether it is relevant or not using Probability Ranking Principle (PRP)? [5]

d5: jan dil pakistan

We need to calculate

$P(R=1/ d_i, q)$ and $P(R=0/ d_i, q)$ for given document d_5

$P(R=1/ d_5, q) = 85.75$
$P(R=0/ d_5, q) = 0.00215$

as $P(R=1/ d_5, q) > P(R=0/ d_5, q)$ the given document is relevant.

< The End>