National University of Computer & Emerging Sciences

Midterm Examination II – Fall 2017(Sol)

Course: IR&TM (CS567) Time Allowed: 1 Hour
Date: October 24, 2017 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: 25 Min] [Marks: 20]**

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 are the two assumptions for which a relevance feedback really works?

1. The user has to have sufficient knowledge to be able to make an initial query which is at least somewhere close to the documents they desire.
2. The relevant documents are similar to each other's; these documents have high overlapping terms. Ideally the term distribution is similar to relevant examples marked by the user.

b. Set oriented evaluation is not good for ranked retrieval. Justify.

In ranked retrieval the result-set is already ordered in term of relevance from the query. The set-oriented evaluation does not suit this kind of order list. We need to assigns more weights to higher order ranked results in evaluation measure.

c. Average precision cannot discriminate a rank order search results with graded relevance. Explain.

Average precision is average of the precision level values at different order of the results. Consider two precision value lists L1: 1,2,3,4, and L2: 4,3,2,1 they produced average precision of 2.5 in each case. Therefore, it cannot discriminate the ordering of precision value in the list.

d. What is the value of nDCG for a perfect ranking algorithm?

nDCG = DCG/ IDCG is a ratio between discount cumulative gain and ideal discount cumulative gain. For a perfect ranking DCG=IDCG so nDCG would be 1.

e. Give three reasons why relevance feedback has been little use in web search.

1. The web users are interested in high precision of the top few results while relevance feedback mainly used to increase recall.
2. The results of relevance feedback sometimes may surprise users and it will be hard to explain the reason of such results.
3. It complicates the search interface and slows down the interactions.

f. What does the assumption “a query term is equally likely to be present or absent from a randomly pick relevant document” in probabilistic IR signify?

With this assumption we have the probability that a query term appears in a relevant document is P (t) =0.5, and a query term absent from relevant document is also Q (t) =0.5 where Q (t) = 1-P (t) [only for relevant collection against a query] which is practical for 1-P (t) and P (t) cancel out each other and thus simplify the expression for probabilistic IR

g. Every term in the document is a random variable. What does this assumption signify in probability ranking principle?

Every term is coming from a sample space Σ* over a finite symbol set λ. The probabilities are computed using a random variable x=t for a term.

h. Define what do we mean by Odd of an event. How it is related to probability ranking?

Odd of an event E is a ratio define as Odd(E)= P(E)/P(~E), the odd values are monotonic and odd of relevance is rank similar to probability of relevance. Rather than estimating this probability directly, because we are interested only in the ranking of documents, we work with odd of relevance which are easier to compute and which give the same ordering of documents.

i. How term independence assumptions are incorporated in both vector space model and probabilistic information retrieval?

In Vector Space Model (VSM), every term is considered as a different dimension hence there is no relationship between terms. They are independent. While in probabilistic information retrieval, we assume that the probability of each term is independent of other term, hence simplify the computation required for probability of relevance.

j. What are the differences between standard vector space tf-idf weighting and the BIM probabilistic retrieval model (in the case where no document relevance information is available)?

When no relevance is given we estimate the probability of a term belong to non-relevant document is u(t)= log (N-df_t / df_t) where df_t is the document frequency of term t and as most documents are non-relevant we have u(t) = log (N/df_t) which is similar to tf*idf factor of vector space model. Hence the two are similar.

Question No. 2 **[Time: 15 Min] [Marks: 10]**

Consider the following given document-collection and a query.

d1: virus microscopic organism
d2: virus infects cell organism
d3: virus infects computers
d4: tiny virus security
q: virus tiny organism

From the system d1 and d2 are relevant to this query and d3 and d4 are not. Using the probabilistic model for IR with given relevance, rank these documents using probability ranking principle. Show all intermediates steps of calculations. Compute whether the document **“d5: virus computer virus”** is relevant or not.

Words	**cell**	**computers**	**infects**	**microscopic**	**organism**	**security**	**tiny**	**virus**
$N_1(W)$	1	0	1	1	2	0	0	2
	0.5	0.1667	0.5	0.5	0.833	0.1667	0.1667	0.833
$N_0(W)$	0	1	1	0	0	1	1	2
	0.1667	0.5	0.5	0.1667	0.1667	0.5	0.5	0.833

Now, we want to rank all documents with the probability of relevance with the query. P(R=1/ di,q) for i=1,2,3, and 4.

RSV for d1= P(R=1/ d1,q) $=^{rank}$ (0.833/0.1667)X(0.833/0.1667)X0.5 = 12.48

RSV for d2= P(R=1/ d2,q) $=^{rank}$ (0.833/0.1667)X(0.833/0.1667)X0.5 = 12.48

RSV for d3= P(R=1/ d3,q) $=^{rank}$ (0.833/0.1667)X(0.1667/0.833)X0.5 = 0.5

RSV for d4= P(R=1/ d4,q) $=^{rank}$ (0.1667/0.833 x 0.5/0.5) X(0.833/0.1667 X 0.1667/0.833)X0.5 = 0.1

Hence ranking will be either d1, d2, d3 and d4 or d2, d1, d3 and d4.

Now, knowing the class of d5;

P(R=1/ d5,q) = ((0.833/0.1667)X(0.1667/0.833))X0.5 = 0.5

P(R=0/ d5,q) = ((0.1667/0.833)X (0.833/0.1667))X0.5 = 0.5

P(R=1/ d5,q) is not greater than P(R=0/ d5,q) hence d5 is non-relevant.

Question No. 3 [Time: 15 Min] [Marks: 10]

Suppose that a user's initial query is q= w1 w3 w2 and IR systems return four documents. User selected d1= w2 w3 w4 and d4= w1 w3 w4 w1 as relevant. While d2= w3 w3 w4 w5 and 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.

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

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

Using the given equation, we will get,

$q_m = \alpha * < 1, 1, 1, 0, 0> + \beta * 1/2 * \{ < 0, 1, 1, 1, 0> + < 2, 0, 1, 1, 0>\} - \gamma * 1/2 * \{< 0, 0, 2, 1, 1> + < 0, 1, 1, 1, 1>\}$

$q_m = \alpha * < 1, 1, 1, 0, 0> + \beta * 1/2 * \{ < 2, 1, 2, 2, 0> \} - \gamma * 1/2 * \{< 0, 1, 3, 2, 2> \}$

$q_m = \alpha * < 1, 1, 1, 0, 0> + \beta * <1,1/2,1,1,0> - \gamma * \{< 0, 1/2, 3/2, 1, 1> \}$

$q_m = < \alpha + \beta , \alpha + \beta/2 - \gamma/2 , \alpha + \beta - 3/2 \gamma, \beta - \gamma, - \gamma>$

We need to put zero on all the dimensions where we have identifiable negative values:

$q_m = < \alpha + \beta , \alpha + \beta/2 - \gamma/2 , \alpha + \beta - 3/2 \gamma, \beta - \gamma, 0>$