National University of Computer & Emerging Sciences

Midterm Examination II – Fall 2016-sol

Course: IR&TM (CS567) Time Allowed: 1 Hour
Date: November 01, 2016 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 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. Why relevance feedback mechanism is not popular among users? Explain.

Relevance feedback is hard to explain to an average user.

Relevance feedback is an extra computational phase in IR cycle, which seeks implicit or explicit judgment about the retrieved documents against a given query and thus delay the query response to a user.

Relevance feedback generally increases recall while an average user interested in high precision.

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.

d. In Rocchio algorithm what will be the condition (values of α, β and γ) for which original query ($\mathbf{q_o}$) is more close to centroid of relevant documents than modified query ($\mathbf{q_m}$).

The original query ($\mathbf{q_o}$) is more close to centroid of relevant documents than modified query ($\mathbf{q_m}$), if the value β is very small and γ is very large and we keep α=1.

e. What is probability ranking principle? Explain.

Let $R_{d,q}$ be an indicator random variable that says whether d is relevant with respect to a given query q. That is, it takes on a value of 1 when the document is relevant and 0 otherwise. The obvious order in which to present documents to the user is to rank documents by their estimated probability of relevance with respect to the information need: P(R = 1|d, q). This is the basis of the Probability Ranking Principle (PRP).

f. In Binary Independence Model (BIM) what does P(R=1/ x, q) + P(R=0/ x, q) =1 assumption represent?

The assumption that a document "d" represented in vector space of terms as "x" is either belong to set of relevant document or set of non-relevant document given a fixed query. [A document is either relevant or non-relevant to a query]

g. In a language model that uses uni-gram and bi-gram probabilities of features how can we calculate the probability of phrase "t_1 t_2 t_3 t_4" respectively. Only probabilities for calculating P (t_1 t_2 t_3 t_4) are required?

The simplest form of language model simply throws away all conditioning context, and estimates each term independently. Such a model is called a uni-gram language model: $P(t_1t_2t_3t_4) = P(t_1)P(t_2)P(t_3)P(t_4)$

bi-gram is a more complex model than uni-gram. This model condition the term based on the previous term hence keep track of order of terms in a limited sense.

P(t1t2t3t4) = P(t1)P(t2|t1)P(t3|t2)P(t4|t3)

h. What is "query expansion"?

Query expansion is an autonomous process of reformulating a seed query (q_o) to improve retrieval performance in information retrieval systems. It is generally perform to bridge the gap between user information need and the posed query. The process usually involves evaluating a user's input (query) to finds synonyms of query terms, morphological forms of the terms, and fixing spelling errors automatically, also re-weighting the terms in the given query to get more relevant documents to a given query. The query expansion add more terms to the query (q_o) to expand the result-set.

i. What does the assumption "a query term is equally likely to be present or absent from a randomly pick relevant document" in BIM 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 BIM.

j. What does the assumption "if a term is not in a query, it is equally likely to occur in relevant and non-relevant collection" in BIM signify?

With this assumption all non-query terms get equal likely value for both relevant and non-relevant sub-collections hence cancel out each other and only the query terms that appears in documents are used for actual calculation in BIM expression.

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

Suppose we have a collection that consists of the 4 documents given below:

D1: good bye see you
D2: good to see you
D3: good morning
D4: good afternoon

Build a query likelihood language model for this document collection. Assume a mixture model between the documents and the collection, with weighting scheme λ=0.5. Let the maximum likelihood estimation (mle) is used to estimate as unigram features. Calculate the model probabilities of the query="good bye to you" and use those probabilities to rank the documents returned by the query.

Doc.	afternoon	bye	good	morning	see	to	you
$M_{(d1)}$	0	1/4	1/4	0	1/4	0	1/4
$M_{(d2)}$	0	0	1/4	0	1/4	1/4	1/4
$M_{(d3)}$	0	0	1/2	1/2	0	0	0
$M_{(d4)}$	1/2	0	1/2	0	0	0	0
$M_{(collection)}$	1/12	1/12	4/12	1/12	2/12	1/12	2/12

λ= ½
q= "good bye to you"

terms	D1	D2	D3	D4
good	7/24	7/24	5/12	5/12
bye	1/6	1/24	1/24	1/24
to	1/24	1/6	1/24	1/24
you	5/24	5/24	1/12	1/12

$P(q/ M_{(d1)}) = 7/24 \times 1/6 \times 1/24 \times 5/24 = 0.000423$
$P(q/ M_{(d2)}) = 7/24 \times 1/24 \times 1/6 \times 5/24 = 0.000423$
$P(q/ M_{(d3)}) = 5/12 \times 1/24 \times 1/24 \times 1/12 = 0.000603$
$P(q/ M_{(d4)}) = 5/12 \times 1/24 \times 1/24 \times 1/12 = 0.000603$

Ranking as per mle

D1, D2
D3, D4

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

Compare and contrast the following three models for information retrieval.

a. Vector space model
b. Probabilistic model
c. Language model

Vector Space Model	**Probabilistic Model**	**Language Model**
Idea: The document and query are represented in vector space of terms appears in both. Similarity is define as the distance between document and query vectors.	Idea: Address the uncertainty in the query. Try to estimate a query term probability to appear in a relevant document. The documents are presented to user with decreasing probability of terms appears in the relevant documents.	Idea: It is also a variation of probabilistic model, it build a model for every document and try to find the probability that a query is generated with the same model of a document. It is related to the idea that user already have some notion of documents in mind while coming up for the query.
Opportunity: - Based on strong mathematical rigor. - Rank documents based on the similarity of query and documents. - Partial and full matching possible.	Opportunity: - Based on general idea of probability theory. - Rank documents based on their probability of relevance to a query. - Partial and full match possible.	Opportunity: - Based on the specific idea of probabilistic document model that generate the given query. - Rank documents based on the probability of relevance of a document model and query. - Partial and full match possible.
Limitations: - No contextual information in the model - Model tuning based on statistical term weighting - High dimensionality	Limitations: - Strong assumption of term independence - Term with zero probabilities need some kind of smoothing	Limitations: - Strong assumption about probability. - Overcome theoretically between language mismatch of documents and query.