National University of Computer & Emerging Sciences, Karachi

Spring-2023 School of Computing

Midterm Exam-II (Sol)

April 05, 2023, 08:30 AM – 09:30 AM

Course Code: CS4051	**Course Name:** Information Retrieval
Instructor Name / Names: Muhammad Rafi	
Student Roll No:	**Section:**

Instructions:

- Return the question paper.
- Read each question completely before answering it. There are **3 questions** on **2 pages**.
- In case of any ambiguity, you may make assumption. But your assumption should not contradict any statement in the question paper.

Time: 60 minutes **Max: 40 Marks**

Question No. 1	<CLO # 1>	**[Time: 25 Min] [Marks: 10X2]**

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 the purpose of relevance feedback in IR?

Relevance feedback refers to an interactive cycle that helps to improve the retrieval performance based on the relevance judgments provided by a user. The idea of relevance feedback 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 and this extra information is used to determine the relevant documents for the next round of retrieval. This also helps to model human perception about query in a better way.

b. Why the first round of the relevance feedback is always very effective?

Engaging user in the retrieval process helps in determining the query intent very quickly and formulating this relevance feedback into the query process increases recall and hence very effective in first round of retrieval. The second round may not be able to improve any further and hence result of further round seldom show any improvement on recall.

c. What is a search snippet? How it is related to relevance feedback?

A search snippet is a short summary of the retrieved document from a search, which is designed so as to allow the user by reading a small portion of text summary, He can give feedback on the relevance of the document.

d. Why Precision and Recall cannot be used to evaluate ranked retrieval systems?

Precision and recall are set-based measures. That is, they evaluate the quality of an unordered set of retrieved documents. To evaluate ranked lists, precision is recall cannot distinguish between different order as both are orthogonal to each other.

e. What is query expansion? When it is useful?

Query expansion (QE) is a process in Information Retrieval which consists of selecting and adding terms to the user's query with the goal of minimizing query-document mismatch and thereby improving retrieval performance. Query expansion is useful when the user query is very ambiguous and there are multiple senses or intent for the given query.

f. What do we mean by Odd of an event? How it is related to ranking of a probability function?

The odds of an event are the ratio of the probability of an event to the probability of its complement. In other words, it is the ratio of favorable outcomes to unfavorable outcomes. In probabilities information retrieval in particular, we can rank documents by their odds of relevance (as the odds of relevance is monotonic with the probability of relevance).

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. The assumption signify that are terms in a language are equal likely and hence it is treated as a random variable.

h. What is the problem with Cumulative Gain(CG) as an evaluation function?

The problem with CG is that it does not take into consideration the rank of the result set when determining the usefulness of a result set. Hence the two results with different ordering with same relevance would have the same score for CG and hence does not cater any difference in ordering of the result set.

i. List any two limitations of the Normalized Discount Cumulative Gain (NDGC)?

1. NDCG metric does not penalize for bad documents in the result.
2. NDCG does not penalize for missing documents in the result.

j. 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.

Question No. 2 **<CLO # 2>** **[Time: 15 Min] [Marks: 10]**

a. Describe at least 2 differences between vector space relevance feedback and probabilistic relevance feedback. [2.5]

(i) In case of the probabilistic (pseudo) relevance feedback, an initial guess is to be done for the relevant document set under the assumption that all query terms give same values of document relevance. (i.e. pi is constant) whereas the vector space feedback system doesn't require any initial assumption or guess as the relevant document set can be computed over the collection and the query.

(ii) The vector space feedback system should ideally perform better by giving improved results in less number of iterations than the probabilistic feedback system due to the assumptions and guesses.

b. Explain Query Likelihood Model for information retrieval. [2.5]

The query likelihood model is a language model used in information retrieval. A language model is constructed for each document in the collection. It is then possible to rank each document by the probability of specific documents given a query. This is interpreted as being the likelihood of a document being relevant given a query. As the same random process is used to generate the query as the process used for a given document.

c. What are the drawbacks of N-gram language model for IR? [2.5]

There are two main drawbacks in N-gram language model (i) data sparsity and (ii) Out-of-Vocabulary term (none-of the feature of training set available in test set). Data sparsity, which means that some n-grams may not occur frequently or at all in the training data, resulting in low or zero probabilities. This can lead to inaccurate or incomplete language models that fail to capture the diversity and variability of natural language. On the other hand, out of vocabulary is also a problem in effectively classifying the test cases.

d. Give at least 2 assumptions of the Binary Independence Model (BIM) for IR? [2.5]

Binary Independence Model (BIM) for IR has the following two string assumptions:

1. The Binary Independence Assumption is that documents are binary vectors. That is, only the presence or absence of terms in documents are recorded.

2. Terms are independently distributed in the set of relevant documents and they are also independently distributed in the set of irrelevant documents.

Question No. 3	**<CLO # 2>**	**[Time: 15 Min] [Marks: 10]**

Consider the following four documents and a query.

Doc 1: click go the shears boys click click click
Doc 2: click click
Doc 3: metal here
Doc 4: metal shears click here
Query: click shears

Build a query likelihood language model for this document collection. Assume a mixture model between the documents and the collection, with both weighted at 0.5. The maximum likelihood estimation (mle) is used to estimate both as the unigram models. Answer the following questions.

a. Compute the model probabilities table for all documents and the collection. [2.5]

Word/Model	**click**	**go**	**the**	**shears**	**boys**	**metal**	**here**
M(Doc1)	4/8	1/8	1/8	1/8	1/8	0	0
M(Doc2)	2/2	0	0	0	0	0	0
M(Doc3)	0	0	0	0	0	1/2	1/2
M(Doc4)	1/4	0	0	1/4	0	1/4	1/4
M(C)	7/16	1/16	1/16	2/16	1/16	2/16	2/16

b. Compute the query probabilities for the query “click shears” [2.5]

As per the mixture model we will have
P(M(d) | q) = Over all Query terms {(λ * P(t|M_d) + (1− λ) P(t|M_c)}
P(M(Doc1)/q) = {1/ 2 * (1/2 +7/16) + 1/2 * (1/8 + 2/16)} = 0.0585
P(M(Doc2)/q) = { 1/ 2 * (0 +2/16)+ 1/2 * (0 + 2/16) }= 0.0039
P(M(Doc3)/q) = { 1/ 2 * (1 +7/16)+ 1/2 * (0 + 2/16) }= 0.0449
P(M(Doc4)/q) = { 1/ 2 * (1/4 +7/16)+ 1/2 * (1/4 + 2/16) }= 0.0644

c. Give the ranking order for the given four documents. [2.5]

Doc 4, Doc 1, Doc 2, Doc 3

d. Why Doc 3 is getting a non-zero probability? Explain [2.5]

It is due to the smoothing method and mixture model between document and query. It is evident that the term that does not appear in the document get some of the score through smoothing and mixture model.

BEST OF LUCK