National University of Computer and Emerging Sciences

Information Retrieval (CS4051)

Date: April 03, 2024

Course Instructor

Muhammad Rafi, Basit Ali

Sessional-II Exam

Total Time: 1 Hour

Total Marks: 40

Total Questions: **03**

Semester: Spring-2024

Campus: Karachi

Department: AI & DS

_____ _____ _____ _____

Student Name Roll No Section Student Signature

_____ _____

Vetted by Vetter Signature

CLO # 1: < Understand the basic concepts and techniques in Information Retrieval.>

Q1. Answer the following questions to the point. **[Time: 30 mins] [Marks: 2x10]**

a. In vector space model, is the order in which the terms occur important for retrieval?

False. The vector space model does not consider order of the terms in a document. It assumes all terms are independent and orthogonal to each other.

b. What do we mean by Champion List? How it facilitates quick relevant scoring of documents?

The idea of champion lists (sometimes also called fancy lists or top docs) is to precompute, for each term **t** in the dictionary, the set of the documents **r** with the highest weights for **w**. The value of **r** and **w**, are chosen in advance. This is to reduce the computation to only these documents that may give you high relevance value to the query.

c. When can IDF value of a term be zero? Explain?

Inverse Document Frequency (IDF) is a weight indicating how commonly a word is used. The more frequent its usage across documents, the lower its score. The lower the score, the less important the word becomes. If a word occurs in every document the idf score will be log(N/df), where df is the document frequency of the term. It will be log (1) =0

d. Compare Local and Global query expansion methods.

Local Query Expansion	Global Query Expansion
- The terms added in query expansion are based on "local" information in the result list retrieved earlier for the user query and feedback on it. - With local expansion the results set more or less reorganized its ranking to present to a user.	- The terms added in query expansion are often based on "global" information that is not query-specific. Like external Thesaurus, user profile, search log etc. - With global expansion user may get new documents in the results set.

e. Differentiate between Direct and Indirect relevance feedback.

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.

f. What are the three important components of information retrieval system's evaluation?

To measure ad hoc information retrieval effectiveness in an standard way, we need a test-set collection consisting of three things:

1. A document collection
2. A test suite of information needs, expressible as queries
3. A set of relevance judgments, standardly a binary assessment of either relevant or non-relevant for each query-document pair.

g. Why might only using one non-relevant document be more effective than using several?

Negative feedback is very hard to implement in a retrieval system. In any collection there are several classes of documents that do not relate to a given query. Hence there can be many confusing vectors for such feedback, converging it to a relevant space is hard. Hence only one non-relevant document is more effective in determining the real intent of the query.

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

The odds are defined as the probability that the event will occur divided by the probability that the event will not occur. In probability ranking principle, documents are ranked in the odd of relevancy vs. non-relevance, magically it has the same ordering as the decreasing order of the probability of relevance for a ranked document list.

i. Do you agree with statement "In order to improve precision, one has to sacrifice recall in the retrieval results"? Justify your answer.

It is false. Generally, in the context of ranked retrieval if a ranking is perfect. There is no need to trade-off between precision and recall.

j. Explain what Normalized Discount Cumulative Gain (NDCG) is.

Normalized Discount Cumulative Gain (NDCG) is a commonly used metric in information retrieval, designed to measure the effectiveness of ranking models by assessing the quality of an ordered list of results or predictions. Both the relevance of each result and its position in the list contribute to the NDCG calculation. It compares rankings to an ideal order where all relevant items are at the top of the list. It uses a graded relevance scale of documents for computation.

CLO # 2: <Understanding how IR Model works >

Q2. Consider the following documents in a collection. **[Time: 15 mins] [Marks: 2x5]**

Doc 1: w1 w2 w4 w5
Doc 2: w1 w3 w3 w6
Doc 3: w2 w4 w5
Doc 4: w2 w4 w1 w5

a. Using term frequency TF= tf(d,f) and inverse document frequency using the idf = log(N/df), (N=4 and df= document frequency of term t.) give all 4 unit vectors of given document set.

Words	tf(d1)	tf(d2)	tf(d3)	tf(d4)	df	idf	q
w1	1	1	0	1	3	0.124939	0
w2	1	0	1	1	3	0.124939	1
w3	0	2	0	0	1	0.60206	1
w4	1	0	1	1	3	0.124939	0
w5	1	0	1	1	3	0.124939	1
w6	0	1	0	0	1	0.60206	0

Words	TF*IDF(d1)	TF*IDF(d2)	TF*IDF(d3)	TF*IDF(d4)	q
w1	0.12493874	0.124938737	0	0.1249387	0
w2	0.12493874	0	0.12493874	0.1249387	0
w3	0	1.204119983	0	0	1
w4	0.12493874	0	0.12493874	0.1249387	1
w5	0.12493874	0	0.12493874	0.1249387	1
w6	0	0.602059991	0	0	0

b. Consider the query q=w3 w4 w5, give the rank order of all 4 documents, using cosine similarity between unit document vectors and query unit vector with simple tf values.

CLO # 3: < Appreciate the importance of data structures, indexes and retrieval efficiency >

Q3. Answer the following questions. **[Time: 15 mins] [Marks: 2x5]**

a. Consider the following documents marked as relevant (R) and Non-relevant (N) for the query vector **q =< 0.5, 0.2,0.1, 0.3,0.4>**

D1 =< 0.2,0.3,0.1,0.2,0.2> - > R
D2 =<0.1,0.1,0.1,0.2,0.3> -> R
D3= <0,0.2,0.2,0.3,0.7> -> N

Conisdering general Rocchio's Algorithm

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

Find the modified query vector after using relevance feedback via Rocchio's approach. Assume α=1; β=0.75 and γ=0.25 [5]

qm = (1) * < 0.5, 0.2, 0.1, 0.3, 0.4> + (0.75) *(1/2) * {< 0.2,0.3,0.1,0.2,0.2> + <0.1,0.1,0.1,0.2,0.3>} – (0.25) * {<0,0.2,0.2,0.3,0.7>}

qm= <0.5,0.2,0.1,0.3,0.4> + <0.225,0.3,0.15,0.3,0.375> - <0,0.05,0.05,0.075,0.175>

qm= <0.725, 0.45, 0.2, 0.425,0.6>

b. The following list of R's and N's represents relevant (R) and non-relevant (N) returned documents in a ranked list of 08 documents retrieved in response to a query from a collection of 1000 documents. The top of the ranked list (the document that the system thinks is most likely to be relevant) is on the left of the list. This list shows 5 relevant documents. Assume that there are 20 relevant documents in total in the collection.

R; R; N; N; R; R; N; R

i. Compute Precision and Recall of the system

Precision = # of relevantly retrieved / total retrieved = 5/8 =

Recall = # of relevantly retrieved / total relevant = 5/20 = 0.25

ii. Average Precision of the system

Average Precision = 1/5 * (1/1+2/2+3/5+4/6+5/8) = 3.8916 / 5 = 0.7783

< The End.>