IT458: Information Retrieval

End Semester Examination

Name: Chinmayi E. Ramakrishna

Roll No: 18ITT113 Date: 18.11.2021

Q1.

(a)

Algorithm A		Algorithm B	
① d9	3	① d1	1
② d2	0	② d31	2
③ d4	0	③ d22	0
④ d5	0	④ d3	4
⑤ d13	4	⑤ d6	2
⑥ d6	2	⑥ d15	0
⑦ d3	4		

① $CG = 13$

② $DCG = \sum_{i=1}^{7} \frac{2^{rel_i} - 1}{\log_2(i+1)}$

$= \frac{2^3-1}{\log(2)} + \frac{2^0-1}{\log(3)} + \frac{2^0-1}{\log(4)} + \frac{2^0-1}{\log(5)} + \frac{2^4-1}{\log(6)} + \frac{2^2-1}{\log(7)} + \frac{2^4-1}{\log(8)}$

① $CG = 9$

② $DCG = \sum_{i=1}^{6} \frac{2^{rel_i} - 1}{\log_2(i+1)}$

$DCG = \frac{2^1-1}{\log(2)} + \frac{2^2-1}{\log(3)} + \frac{2^4-1}{\log(5)} + \frac{2^2-1}{\log(6)}$

Chinmayi (181IT113)

DCG = 62.6894 DCG = 34.925

③ $IDCG = \frac{2^4-1}{\log(2)} + \frac{2^4-1}{\log(3)} + \frac{2^3-1}{\log(4)} + \frac{2^2-1}{\log(5)} = 97.186$

$IDCG = \frac{2^4-1}{\log(2)} + \frac{2^2-1}{\log(3)} + \frac{2^2-1}{\log(4)} + \frac{2^1-1}{\log(5)} = 62.53$

④ $nDCG = \frac{DCG}{IDCG} = \frac{62.68}{97.18} = 0.64$

$nDCG = \frac{34.92}{62.53} = 0.558$

According to grading based DCG metric (Discounted Cumulative Grading)

→ Algorithm A has better ranking than B.

(b) Discounted Cumulative Grading is the metric and algorithm used since only categories of relevant and irrelevant documents are given.

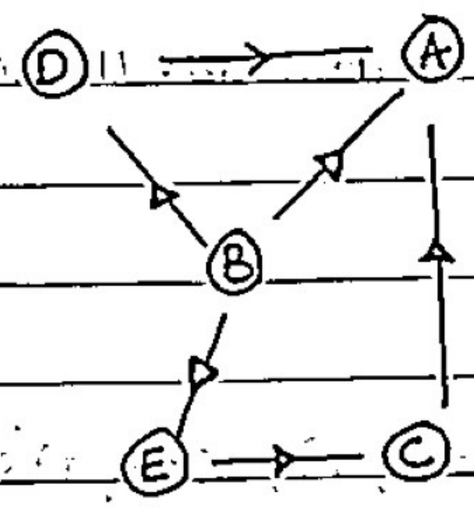
Each algorithm has retrieved documents as follows

4/7 retrieved documents are relevant in Algorithm A

4/6 retrieved documents are relevant in Algorithm B.

Chinmayi (181IT113)

Q2)



$$PR(u) = (1-d) \times \frac{1}{N} + d \times \sum \frac{PR(v)}{C(v)}$$

$d = 0.85$

$$PR(A) = (1-0.85) \times \frac{1}{5} + 0.85 \times \left(\frac{PR(B)}{3} + \frac{PR(D)}{1} + \frac{PR(C)}{1} \right) = 0.124$$

$$PR(B) = (1-0.85) \times \frac{1}{5} + 0.85 \times (0) = 0.03$$

$$PR(C) = (1-0.85) \times \frac{1}{5} + 0.85 \times \left(\frac{PR(E)}{1} \right) = (0.03) + 0.85 \times 0.03) = 0.0627$$

$$PR(D) = (1-0.85) \times \frac{1}{5} + 0.85 \times \left(\frac{PR(B)}{3} \right) = 0.0385$$

$$PR(E) = 0.15 \times \frac{1}{5} + 0.85 \times \left(\frac{PR(B)}{3} \right) = 0.0385$$

The page rank of each web page is as follows:

$PR(A) = 0.124$

$PR(B) = 0.03$ (since no page points to B)

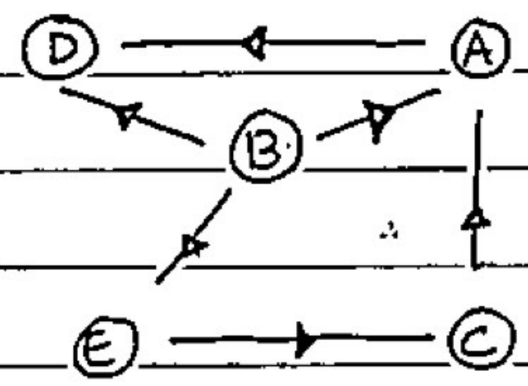
$PR(C) = 0.0627$

$PR(D) = PR(E) = 0.0385$

3

Chinmayi (18IT113)

If A points to D and say the graph becomes



$$PR(A) = 0.03 + 0.85 \times \left(\frac{PR(B)}{1} + \frac{PR(C)}{1}\right) = 0.0917$$

$$PR(B) = 0.03$$

$$PR(C) = 0.03 + 0.85\left(\frac{PR(E)}{1}\right) = 0.0627$$

$$PR(D) = 0.03 + 0.85 \times \left(\frac{PR(B)}{3} + \frac{PR(A)}{1}\right) = 0.116445$$

$$PR(E) = 0.03 + 0.85 \times \left(\frac{PR(B)}{3}\right) = 0.0385$$

By changing the link that went from D to A to A to D, the page ranks have changed to following

$PR(A) = 0.0917$

$PR(B) = 0.03$

$PR(C) = 0.0627$

$PR(D) = 0.116$

$PR(E) = 0.0385$

4

Chinmayi (181IT113)

page ranks of A and D have been changed

original graph	changed graph
① A	① D
② C	② A
③ D	③ C
④ E	④ E
⑤ B	⑤ B

Q3.

1. Identifying questions to show to new users.

One method is to use Social Data Mining.
A new user might have used a similar web page like Stack Overflow in the past. So data from those websites can help in recommending similar questions to the new user through user profiling.

Another method would be through collaborative filtering.
For example, if a student studying Data Science has registered to Stack Overflow, then questions of other Data Science students can be recommended.

5

Chinmayi (181IT113)

2. Capturing the importance of a particular user in the community.

Collaborative Filtering method is suggested (CF)

The idea of CF or community based model is to bring together like minded people. Nearest neighbour matching can be done to understand the taste and hence capture the importance of a particular user. A user whose content or taste is top ranked can be considered important in the community.

3. Query suggestion or auto completion

Language models like unigram, N-gram can be used to predict the next words in the query. Also, the user profile logs can be helpful in understanding the query intent and hence help in query suggestion.

For example, saved passwords or if a person frequently searches for famous or top 10 restaurants, the IR model can suggest terms in the query.

Chinmayi (181IT113)

Q4.

Requirements in the IR model :-

1) Generate a list of keywords for constructing a Research Interests profile
2) Use a corpus of faculty pulications
3) Capture similarity of words to identify similar research areas
4) The corpus is a growing data

General formalisms of an IR model

Documents in the corpus
Query language (boolean or natural language)
Ranking
Feedback to improve the model

Vector Space Model seems best fit for the requirements.

A vector space model provides special representation of documents.

A list of index terms can be extracted from the documents after lemmatization. Lemmatization helps in understanding the core idea of a word/phrase.

7

chinmayi (181IT113)

A similarity score like cosine similarity score can be used to show the similarity between two publications.

The reasons for its importance/relevance over other models are :-

① It can have spatial representation of documents which is not shown by boolean model, fuzzy set model

② It can provide a similarity score between 0 and 1 unlike boolean model which only goes for either 0 or 1.

③ other models like probabilistic models give more value to the overall ranking of the documents rather than the similarities or categorisation of similar publications.

④ Vector space model can use nearest neighbour matching to categorise similar publications.

⑤ New documents can be accomodated by adding their index terms and considering similarity scores.

8