NUI Galway
OÉ Gaillimh

Autumn Examinations 2021-2022

Course Instance Code(s)	4BCT, 1CSD1, 1CSD2, 4BS2
Exam(s)	4th B.Sc. Computer Science and IT M.Sc. Computer Science (Data Analytics) M.Sc. Computer Science (Artificial Intelligence) B.Sc. (Hons)
Module Code(s)	CT4100
Module(s)	Information Retrieval
Paper No.	1
External Examiner(s)	Dr. Ramona Trestian, Dr. John Woodword
Internal Examiner(s)	Professor M. Madden *Dr. C. O'Riordan

Instructions: Answer any 3 questions. All questions carry equal marks

Duration	2 hours
No. of Pages	3
Discipline(s)	Computer Science
Course Co-ordinator(s)	Dr. Colm O'Riordan, Dr. Frank Glavin, Dr. Mattias Nickels

Requirements:

Release in Exam Venue Yes

CT4100 Information Retrieval

Question 1 (25 marks)

(a) Recommender systems are used to generate recommendations for users on unseen items. Collaborative filtering is one such approach. Explain, in your own words, the main stages of collaborative filtering and illustrate how this approach can be used to generate a recommendation. (10)

(b) Suggest approaches to evaluate the usefulness of a collaborative filtering system. Discuss any limitations of these approaches (8)

(c) Outline a suitable approach to incorporating content information about items into a recommender system. (7)

Question 2 (25 marks)

(a) Precision and recall are often used to evaluate the performance of an IR system. Explain these terms. Given a ranked answer set and evaluation judgements, explain how you would generate a precision-recall graph. (8)

(b) Outline any limitations associated with using precision and recall and suggest alternative evaluation measures. (7)

(c) Learning mechanisms has been used successfully in information retrieval. Using an approach of our choice, suggest a learning mechanism to identify good weights to apply to terms in documents. Discuss any limitations of your approach (10)

(1A) 3 main steps in collaborative filtering:

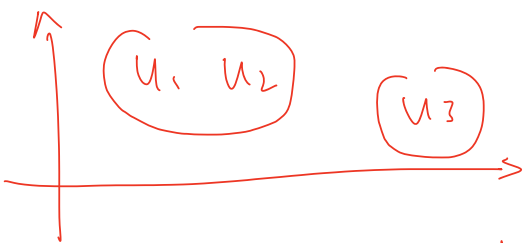
(1) Find how similar each user to every other user.

ratings / user	I_1	I_2	I_3	I_4
U_1	0	10	5	0
U_2	1	9	3	8
U_3	0	0	1	2

→ U_1 & U_2 is similar

★ with Cosine similarity / pearson correlation

(2) form clusters of users who are similar



★ with correlation thresholding / best-n correlation.

(3) In each cluster, make recommendations based on what other users in the group have rated which not yet rated by active user.

→ recommend I_4 to U_1 as U_2 has given high ratings to I_4.

★ with computed weighted average with user ratings. / compute weighted mean with neighbourhood ratings.

(1b) Evaluation metrics for CF.

(1) Coverage - → ability of system to provide a recommend action on a given item.

(2) accuracy : → correctness of the recommendations generated by the system.

Approaches to evaluate:
- MAE, RMSE → to measure accuracy.
  - error as percentage
  - contributions of the absolute errors between the prediction & the true value.

downside : The accuracy measurement might not be useful to users as they might asking for movies that are unfamiliar with but not old favorites they do not likely want again.

(1c) · for each ~~of the item~~, extract key properties such as brand, type, functionality, genre and etc.

Then based on the ratings of each user we could suggest new items that have similar characteristics of high ratings to user (there are coverages).

for new user, we could suggest popular items based on the ratings of all users along with item characteristics matching

for users with no similar ratings as others, some method applies.

(2A) · precision → the correctness in returned set $\frac{TP}{TP+FP}$ $\frac{|RA|}{|A|}$ → answer

recall → → the returned percentage across entire relevant set.

$\frac{TP}{TP+FN}$ $\frac{|RA|}{|R|}$ → relevant set

precision-recall graph.

→ should only be generated from ranked list.

Top 10 ranked list :

↳ d1 d2 d3 d4 d5 d6 d7 d8 d9 d10

underlined = relevant doc.

document set = 20

relevant in document = 10

Then we can calculate the precision-recall pairs one by one

1. Considering 1st document only:

precision = $\frac{1}{1}$, recall = $\frac{1}{10}$

2. considering first 2 doc :

precision = $\frac{2}{2}$, recall = $\frac{2}{10}$

3. considering first 3 docs

$p = \frac{2}{3}$, recall $= \frac{2}{10}$

4. considering first 4 docs.

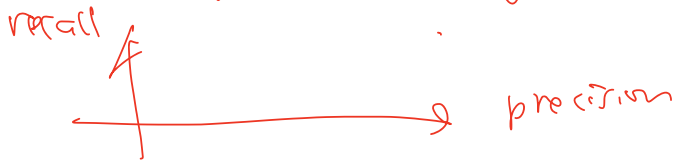
$p = \frac{2}{4}$, $r = \frac{2}{10}$

5. first 5 :

$p = \frac{3}{5}$, $r = \frac{3}{10}$.

then continue until the last answer.

to get 10 (x, y) points to plot.



(2b) - easily achieved recall 100% by returning everything

- precision recall graph - impractical used would be a large list.
- might not be relevant as not measuring the true performance of system by only 1 single query
- alternative (1) average precisions of n queries.
  (2) then take the mean of the average precision for MAP as a single value measurement for

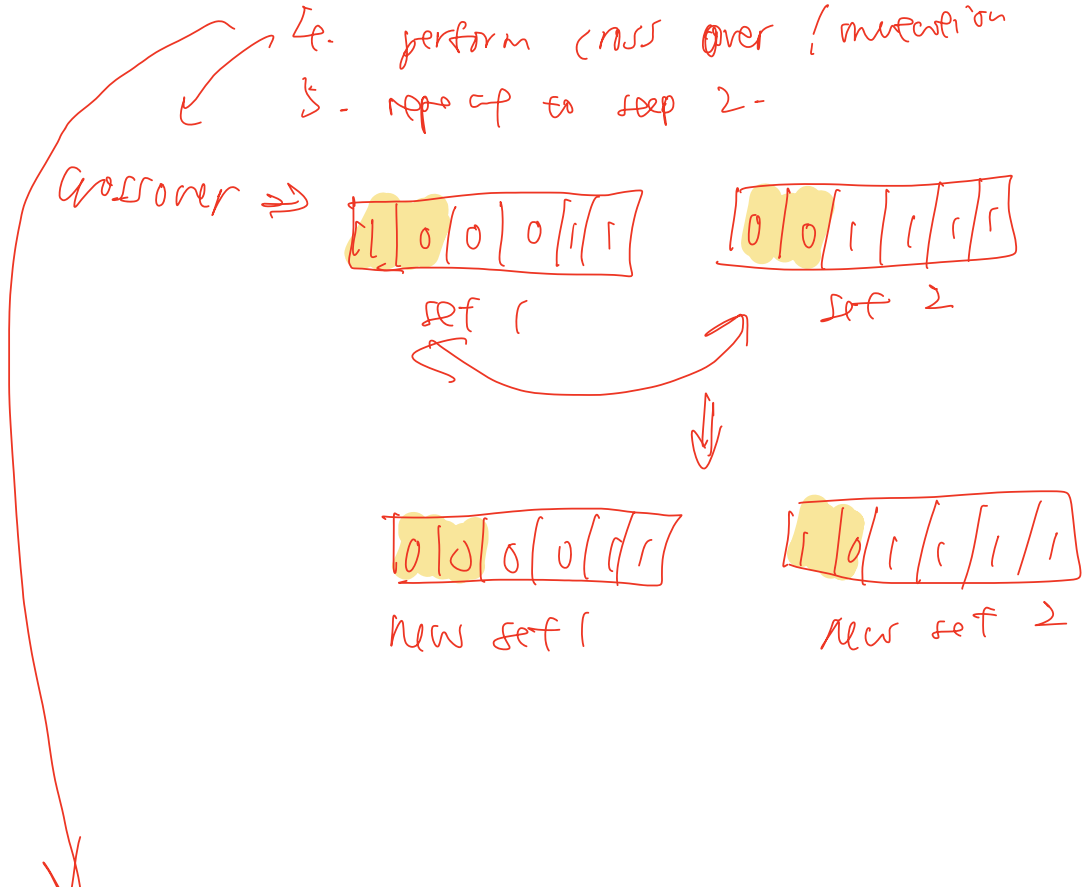
the system.

(3) use f-measure which combines precision & recall

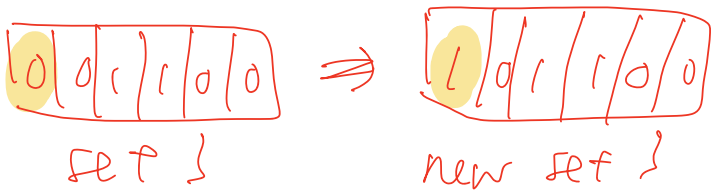
(2c) LR - genetic algorithm

steps :

1. create random population (genotype) with length = size of lexicon + random rules
2. calculate fitness with evaluation function to rescue in map
3. select good population
4. perform cross over / mutation
5. repeat to step 2.



VI

mutation ⇒ 

set } new set }

limitation : ① choice of fitness function

② choice of representation for encoding

③ parameters setting : I - size of population
II - number of generations
III - probability of crossover
IV - probability of mutation

④ definition of selection scheme

Question 3 (25 marks)

(a) Describe, in your own words, with reference to any well-known term weighting scheme, the main constituents of a good weighting scheme. (9)

(b) Local analysis of the returned answer set is often used to select a number of terms to add to the current query. Describe an approach that may be used to select candidate terms to add to the query. Discuss the advantages and potential limitations of your approach. (8)

(c) Suggest an approach to extend local analysis in order to ensure a diverse set of candidate terms. (8)

Question 4 (25 marks)

(a) Discuss approaches to identify the difficulty of a user provided query. Your answer should include features that are used to measure the level of difficulty and you should distinguish between pre-retrieval and post-retrieval approaches. (8)

(b) Given a set of scientific articles (which contain title, abstract, authors, key words, main body of the paper and a bibliography), suggest a suitable approach to measure the similarity between these documents in the collection into useful sub clusters that may be of use in user search tasks. (8)

(c) Given the collection described in (b) and using the similarity measure(s), you have defined, describe an approach to clustering this collection in sub collections that may be of use to users engaging in search tasks. (9)

(3A) - A good weighting scheme should obey the constraints introduced in the axiomatic approach.

↳ for example the BM25 - which:

$$\sum_{t \in Q \cap D} \left(\frac{tf_t^b}{tf_t^b + K_i} \cdot \frac{\log\left(\frac{N - df_t + 0.5}{df_t + 0.5}\right)}{(1-b) + b} \cdot \frac{tf_t^Q}{\frac{dl}{dl_{avg}}} \right)$$

tf idf Query.

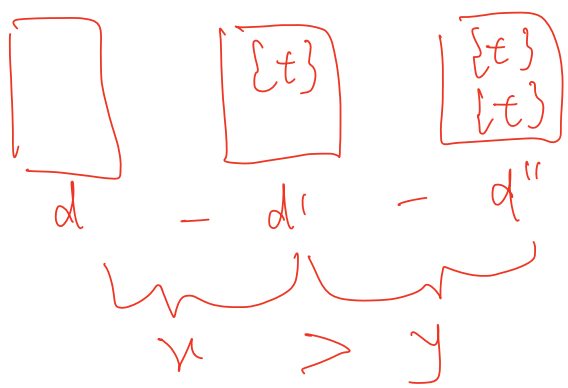
fulfill

1. adding query term increases score of document
2. adding non-query term decreases score of document
3. adding successive occurrence of term to a document increase less score than the addition of successive occurrence.

$$sim(d, q) - sim(d, q)$$

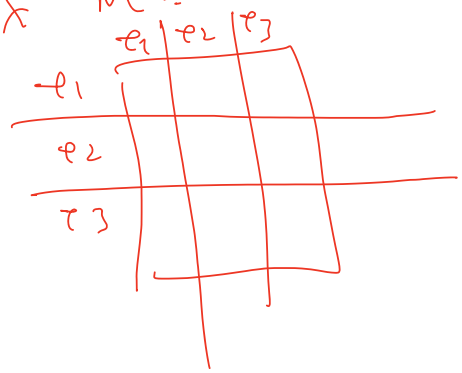
more than

$$sim(d'', q) - sim(d', q)$$



If using a sub linear term-frequency factor component with K_1 as controller. To ensure appearance of non-query term weighted less.

(3B) Association clusters is to find out the correlation of 2 terms by constructing a stylo matrix M :



and formula :

$$M_{ij} = \frac{freq_{i,j}}{freq_i + freq_j - freq_{ij}}$$

then we could select Top N terms that correlates to $term_i$ based on the matrix.

Pros : Cheaper computational cost.

Cons : Less evidence, no difference & no context.

(3C) Clustering terms were selected from 3b.

then select N% from each / top 10 cluster.

(14) - pre-retrieval approaches :

estimate difficulty without running system.

(1) Linguistic approach:

I: To identify if query is ambiguous. (probabilistic parsing.)

(2) statistical approach :

I: Identify the frequency distribution of the query terms in the collection (Idf / Icf).

II: Identify the non-specific terms in queries.

III: Identify term relatedness (mutual information / Jaccard coefficient)

IV: Identify query scope.

↳ how many docs contain the query term.

(measure difference btwn collection language model & query language model)

post retrieval approaches :

explore results from system of query

(1) Clarity: Compare language model induced from result set and language model induced from collection set.

(2) robustness : measure by a deviation of :

I - query : generate a sub-query and measure overlap.

II. document : apply same query in modified version of collection

III : system : search for same query in same collection but different system

(5) Score analysis :

I : measure difficulty based on distribution of values in both results and collection set.

(4b) Measure similarity between articles :

Vector space model (cosine similarity)

↳ can use BM25 to generate weights for each article

(4c) k-means clustering -

↳ use 4b to compute relatedness between each vector. (euclidean distance)