CSIT 6000I Search Engines and Applications
Final Examination, Fall 2019
Released: Dec 13, 2019 1:30pm (HK Time)
Due: Dec 14, 2019 1:30pm (HK Time)

Name: Tomas Sousa Pereira **Student ID: 20667036**

Submission: Submission your answer paper to Canvas. This exam can be completed without programming. If you choose to write programs to verify your results, there is NO NEED to submit your program.

When essay style answers are expected, I give a soft maximum to the number of sentences you can use in your answer.

My estimation is that these questions could be completed in 2.5 hours in a regular exam. Including the time you might spend on reading slides, papers and searching the web, I expect the exam to take less than 5 hours.

---

1. [10] In the traditional relevance feedback formula, document vectors are added or subtracted from the original query to generate a new query for the next round of retrieval. Discuss the disadvantages of this method in terms of efficiency and effectiveness of the feedback query. [max: 5 sentences]

Those operations try to move our query closer to the relevant documents and further away from the non-relevant.

Regarding the efficiency this method has some drawbacks, because we have to count all the relevant and non-relevant documents it will have more time complexity than before we can reduce it by computing the Euclidean distance to the centroid but even this has higher computational cost.

About the effectiveness it gets a bit more complicated, this subtraction often fails to have good results on multi-modal classes. The subset of documents can use different vocabulary pushing two queries away even though they have similar origins.

2. [10] The term-document matrix A is factorized with SVD, and reduced into rank-2 space. Query q is also reduced into the same rank-2 space. Inner product similarity is computed between q and the documents The following table shows the t-d matrix A and query q:

A

	d1	d2	d3	d4	d5	d6	d7
t1	1	1	1	1	1	0	1
t2	0	1	1	1	1	0	1
t3	0	0	0	0	0	1	1
t4	0	0	0	1	0	0	0
t5	0	0	1	0	0	0	0
t6	0	0	0	1	0	0	0
t7	1	0	0	0	1	1	0
t8	0	0	1	1	0	0	0
t9	0	0	0	1	0	0	0

q

	[,1]
t1	0
t2	1
t3	0
t4	0
t5	0
t6	0
t7	1
t8	0
t9	0

After applying SVD to A, we reduce A and q into rank-2 space and compute inner product similarity between q and the documents. Here, I ignore the details of the computation but just show the result, which may be assumed to be correct. The following table shows the similarity scores of the 7 documents:

	[,1]
[1,]	0.173
[2,]	0.086
[3,]	0.044
[4,]	-0.061
[5,]	0.199
[6,]	0.180
[7,]	0.153

The question is: d1 and d2 both match 1 query term with equal weight (highlighted in A for your convenience), so they have the same similarity to q in the original space. However, in the rank-2 space, d2's score is half of d1's score in the rank-2 space; give a plausible explanation of this (one that you believe in) and justify. [max: 8 sentences]

Note: this question does not require you to actually obtain the UDV matrices, although I would not prevent you from doing it if you find it useful in your explanation.

Assuming that we remove all the stops words. Due to the term frequency t1 and t2 can be very general terms with lot of meanings also known as Polysemous. When computing the SVD we only stay with the most important terms, and t7 due to its frequency is more important than t2. Because of this d1 has a higher score than d3.

3. **[5]** Answer the following questions about SVD (one sentence for each part). Suppose $A = UDV^T$,

(a) Which vectors (columns or rows) in the matrices are orthogonal?
U has orthogonal columns. V^T has orthogonal columns.

(b) Which matrix is a diagonal matrix?
D

(c) Which values indicate the importance of the latent factors/indexes?
D

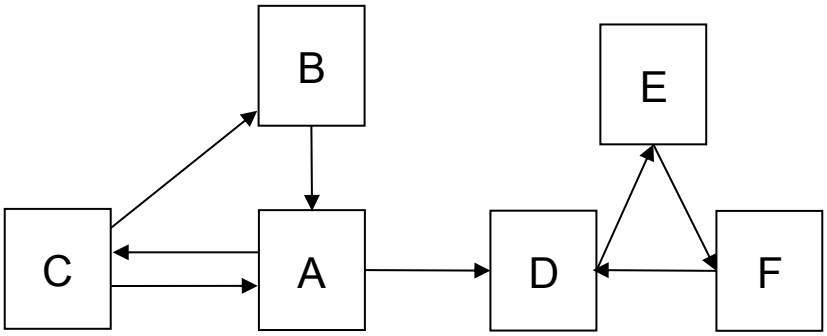
(d) Does SVD produce unique result?

No. D will be unique. But U and V aren't unique, since they are deeply connected i.e. U determines V and you have some freedom to choose U. In the end you have to make sure that $A = UDV^T$,

(e) Do latent factors/indexes correspond to real world concepts.

Latent factors describe a property or a concept that a user or item have. For example, in movies recommendation system a latent factor can be the genre which that the movie belongs to.

4. **[15**] Given the following web graph, compute the PageRank values of the pages by filling in the tables below. The damping factor d=0.8. Apply L1 normalization in each iteration.
Note: In this question, you are NOT required to write program to fill in the table. However, I would not forbid you from writing a program to compute the values, but if you do, you still need to fill in the value of each cell in the table.



(a) [3] List the PageRank formula for each node in the box below:

$$PR(A) = (1 - 0.8) + 0.8\left(PR(B) + \frac{PR(C)}{2}\right)$$
$$PR(B) = (1 - 0.8) + 0.8(\frac{PR(C)}{2})$$
$$PR(C) = (1 - 0.8) + 0.8\left(\frac{PR(A)}{2}\right)$$
$$PR(D) = (1 - 0.8) + 0.8\left(PR(F) + \frac{PR(A)}{2}\right)$$
$$PR(E) = (1 - 0.8) + 0.8\big(PR(D)\big)$$
$$PR(F) = (1 - 0.8) + 0.8\big(PR(E)\big)$$

(b) [6] Fill in the PageRank values, before and after L1 normalization, in each iteration.

Iteration	**0**	**1**	**Normalized**	**2**	**Normalized**	**3**	**Normalized**
PageRank(A)	1/6	0.4	0.2	0.36	0.18	0.36	0.18
PageRank(B)	1/6	0.27	0.13	0.25	0.13	0.26	0.13
PageRank(C)	1/6	0.27	0.13	0.28	0.14	0.27	0.14

Page Rank(D)	1/6	0.4	0.2	0.41	0.21	0.41	0.2
Page Rank(E)	1/6	0.33	0.17	0.36	0.18	0.37	0.18
Page Rank(F)	1/6	0.33	0.17	0.33	0.17	0.34	0.17

(c) [6] Discuss the effect of the rank sink D-E-F on the PageRank values of the nodes when **(i)** there is no teleporting, i.e., d=1, **(ii)** d=0.8, and **(iii)** a link E→B is added to the web graph. [max: 6 sentences totally]

i) When no teleporting after some iterations the PR of both three nodes will reach its maximum, while the other nodes will tend to zero.

ii) Because the dumping factor is 0.8 the rank sink nodes will be attenuated. We are going to minimize the effect of the loop.

iii) In this way the rank sink will be broken, we don't have a loop. The PR of B will increase and PR of F will decrease.

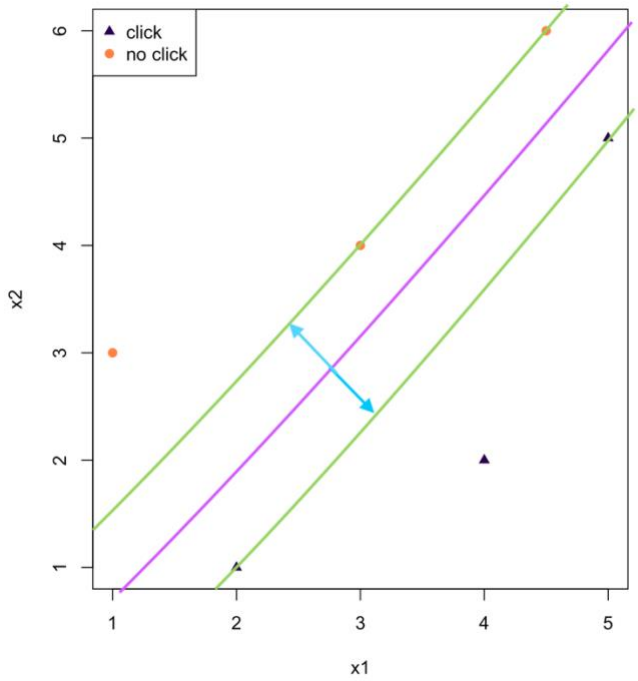
5. [25] A user issues a query q to a search engine and obtains the results shown in the table below. His clicks are shown too. Assume that a click indicates relevancy and a non-click indicates non-relevancy, and the documents have two features X1 and X2.

	X1	X2	click

d1	2.0	1.0	√ (1st)
d2	1.0	3.0	
d3	4.0	2.0	√ (2nd)
d4	3.0	4.0	
d5	5	5	√ (3rd)
d6	4.5	6	

(a) [10]

i) **(**Plot the data points on a two-dimensional graph with X1 as x-axis and X2 as y-axis. Identify by observation the decision (discrimination) function separating the relevant and non-relevant documents, and draw the decision function, the margins and the support vectors on the graph.



ii) Based on intuition and visual inspection of the figure, which term is more important in determining the relevance of the documents? [max: 3 sentences]
X2, observing the figure we can conclude that 66% of the clicks occur if x2<3. On the other hand we cant observe any conclusion about x1.

You can use software to generate the plot and copy-and-paste it into your answer, but hand-drawn with clear labels is enough.

(b) [5] Using the preference mining rules, identify the preference pairs for each of the rules in the table below

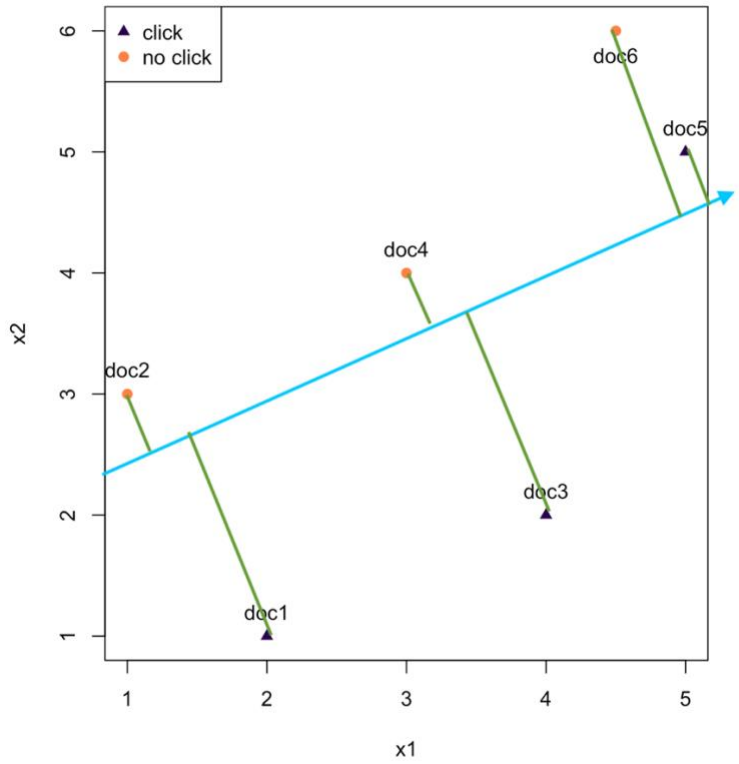
Click > Skip Above	Last Click > Skip Above	Click > Earlier Click	Last Click > Skip Previous	Click > No-Click Next
d2<d3 d2<d5 d4<d5	d2<d5 d4<d5	d1<d3 d1<d5 d3<d5	d2<d3 d4<d5	d1>d2 d3>d4 d5>d6

List the preference pairs that are unique across all rules in the box below.

d1<d3, d1<d5, d3<d5,d1>d2, d3>d4, d5>d6

(c) [10] Plot the same documents on a new graph. Draw a ranking vector, clearly indicating the direction of the vector (vector pointing to direction of high preference). If any preferences cannot be satisfied by your decision function, explain why. [max: 5 sentences]

We can satisfy all of the preferences



6. **[5]** Describe in words the novelty and diversity recommendation performance metrics, and whether or not they conflict with accuracy-based recommendation performance metrics (the improvement of one metric will often cause the other metric to degrade. [max: 6 sentences]

   Diversity is a metric that measures the number of different things recommended by the RS, i.e. how many different movies genres we recommend In our movie list.

   Novelty measures the surprise of a recommended item to the user. The recommendations of our RS should be items which the user wasn't expecting.

   More diversity gives the user more general items but makes creates more chances for the user dislike our recommendations. Novelty surprises the user with new items, but the user can dislike these new items. Is common for both to have a negative impact on the accuracy.

7. **[5]** SVD and Latent Factor Model (LFM, ref: Koren's paper) aim to factorize a matrix into two smaller matrices, the product of which approximates the original matrix. [max: 2 sentences of each part]

   i) Is there any common/shared objectives between SVD and LFM? please explain;

   SVD allows us to find the best features of that matrix, i.e. for a term-doc matrix we will find the most important weight on each doc. LFM allows to map the user-item interaction and map in a lower space, they both try to do the similar things.

   ii) Explain why it is not appropriate to apply SVD to factorize a user-item rating matrix.

   User-item rating matrix has a problem with sparse values SVD is undefined when knowledge about the matrix is incomplete. Some systems try to fill the miss data but this method can distort the data.

8. **[5]** In Koren's paper, regularization is used to control the learnt parameters.

i) What is the purpose of regularization as mentioned by the authors;
Instead of making the systems learn with the previously observed ratings and create the problem of overfitting. The idea of Regularization is to generalize the previous ratings in a way that predicts the future unknown ratings.

ii) What does regularization try to control?
Regularization tries to control the learned parameters whose magnitudes are penalized.

iii) Beyond the raw user-item click action, the authors incorporated two additional factors into their latent factorization model. What are they and explain what they represent [max: 2 sentences for each part]

Implicit feedback allows us to use the user behavior to have better data, we can have a Boolean implicit denoting if the user has implicit preference about that item i.e. we will have a list with items which the user has showed some preference (mouse behavior, history, etc.).
User attributes gives us more information about the user such as gender, age group, income level, etc. this information can impact our RS i.e. if our user is a kid, we should recommend more content for kids.

9. **[5]** In the following rating matrix, ratings are given from 0 to 1 and blanks mean missing/unknown ratings.

	i1	i2	i3	i4	i5	i6	i7	i8	i0
u1									
u2			0.2		0.3		0.3	0.1	0.4
u3	0.4		0.3			0.2	0.4		
u4			0.9		0.4		???		0.5
u5		0.4	0.7		0.6		0.7		0.5

(i) [1] Are there any cold-start users or items in the matrix? Please identify them, if any.
u1 and i4

(ii) [4] Using user-based collaborative filtering, predict the rating of u4 about i7, taking into account user-user similarity and user lenience/toughness into consideration.

$sim(u2, u4) = 0.2 * 0.9 + 0.3 * 0.4 + 0.4 * 0.5 = 0.5$
$sim(u3, u4) = 0.3 * 0.9 = 0.27$
$sim(u5, u4) = 0.7 * 0.9 + 0.5 * 0.5 + 0.6 * 0.4 = 1.12$

$weight(u4, i7) = (0.3 * 0.5 + 0.4 * 0.27 + 0.7 * 1.12)/1.89$

$$avg(u2) = \frac{(0.2 + 0.3 + 0.3 + 0.1 + 0.4)}{5} = 0.26$$
$$avg(u3) = \frac{0.4 + 0.3 + 0.2 + 0.4}{4} = 0.325$$

$$avg(u4) = \frac{0.9 + 0.4 + 0.5}{3} = 0.6$$
$$avg(u5) = \frac{0.4 + 0.7 + 0.6 + 0.7 + 0.5}{5} = 0.58$$
$$delta(u2, i7) = 0.3 - 0.26 = 0.04$$
$$delta(u3, i7) = 0.4 - 0.325 = 0.075$$
$$delta(u5, i7) = 0.7 - 0.48 = 0.12$$
$$weightTough(u4, i7) = 0.6 + (0.5 * 0.04 + 0.27 * 0.075 + 1.12 * 0.12)/1.89 = 0.69$$

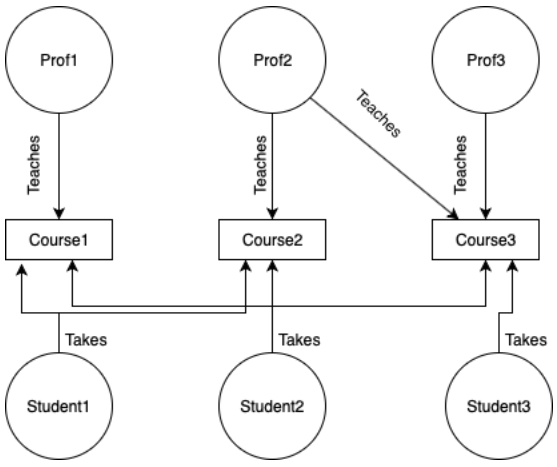
10. **[5]** In the Alibaba/Taobao paper, the authors suggest to include "side information" into the learning framework.

i) List the 3 most significant side information obtained in the paper;
Embedding of an item itself. Shop. Gender. (if we don't count with the item itself Age should be the next one)

ii) What is the problem side information is trying to tackle? [max: two sentences for each part]
Side information is mainly trying to tackle the cold-start and the sparsity problem. Using side information the authors try to obtain accurate embeddings of items with low or even no interaction.

11. **[5]** If a Heterogeneous Information Network (HIN) is designed to capture students' course enrollment information, involving students taking courses taught by professors,

i) Draw a HIN schema and a small instance graph showing 2-3 instances for each node type;



ii) identify two metapaths that can be used for computing the similarity between two students.

Student1 (takes)→ Course2← (teaches) Professor2 (teaches) → Course3←(takes) Student3

Student1 (takes)→ Course1←(takes) Student2 (takes) → Course3 ←(takes) Student3

12. **[5]** If we say that in the physical world, the space is defined by latitude, longitude and altitude, a person is mapped into the space based on her location in the space, and the distance between two persons is defined by Euclidean distance. What can you say about the document world in terms of document space, documents and similarity by revising the first sentence? [max: 5 sentences]

We would have almost 8 billions documents each with at least 3 terms (latitude, longitude and altitude)
Similarity is given be the Euclidean distance, which can also be seen on normal document-term.

With the similarity we could see the people who are closer to each other e.g. family friends.