CSIT 5930 Search Engines and Applications

# Mid-Term Examination, Fall 2020

# Released: Nov 7, 2020 2pm (HK Time)

# Deadline: Nov 7, 2020 4pm (HK Time)

|  |
| --- |
| ***Required:*** *Type in the above statement in the space below:*  *Full Name:* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  *Student ID:* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ *Date:* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

**Submission: Submission your answer paper to Canvas.**

Please read the following clearly:

* You can type in your answer in this Word file and submit a PDF version of it on Canvas; a PDF version of this file is available for your reference.
* This exam can be completed without programming. If you choose to write programs to verify your computational results, the time is at your own expense, and there is NO NEED to submit your program.
* If the questions are unclear or need clarifications, email to [dlee@cse.ust.hk](mailto:dlee@cse.ust.hk), and continue with other questions until you get a reply.
* I will send email to send clarifications, so check your emails frequently.
* If you feel that the question is not clear and information is sufficient for you to answer the questions, you can write down your assumptions for consideration in grading.
* You need to submit your answer file to Canvas before the deadline. To avoid overloading Canvas, I will keep the system open for 5 min after the deadline. However, you cannot modify the file after the deadline, and we will check the last modified time of your submission.

1. **[5]** Refer to the PageRank formula, what are the factor(s) that contribute to the PageRank of a page?

(a) The number of the page’s parents increases

(b) The number of the page’s grandparents increases

(c) The PageRank of the parents increases

1. **[6]** In the vector space model, (a) when stemming has been applied to document terms, explain if it is necessary to apply stemming to the query terms. (b) When stop word removal has been applied to document terms, is stop word removal required on the query terms.

(a) Yes, it is necessary to stem the query terms; otherwise, for the same word, the stemmed and un-stemmed versions will not match.

(b) No, it is desirable but not a requirement. The same documents will be retrieved, albert with different scores. For example, if the query is science and technology, and the documents have all “and” removed, documents containing “science” and “technology” will still match the query in the vector space model. The tricky part is the word “required” in the question.

1. **[6]** Give two reasons for why using links to infer the similarity between two linked pages cannot improve web search quality.

I give three that I can think of:

For web search, there are too many truly relevant documents for a query. Link-based similarity can only improve recall but not necessarily precision. Given the size of the web and the type of users, precision is much more important than recall.

Given the wide diversity in quality, similarity cannot distinguish a page written by a professor from a page written by a high-school student on the same topic (e.g., robots). We need a new dimension to measure authority/quality, and link-based similarity does not help in this case.

Links, at best, only indicate the fact that two pages are related but not necessarily similar in topic, e.g., movies may link to music, but a person searching for movies are primarily interested in movies, not the music.

1. **[6]** Give one advantage and one disadvantage of Clever (the search engine architecture, not HITS) over Google's way of integrating PageRank into keyword-based ranking. Justify your answer.

Advantage: Clever retrieves a set of related pages from a traditional search engine for computing the hub and authority weights. Thus, the hub and authority weights are topical; for PageRank, it is computed on the whole graph, the PageRank of a page on sport may be gained from pages about archeology.

Disadvantages of Clever: Clever computes authority and hub weights in query processing time, whereas Google precomputes the global, static PageRank for all pages and thus have fast query processing.  
  
Disadvantages of Clever: Clever computes authority and hub weights only on a small subgraph, so the weights computed are not as representative as if all pages related to the queries are used.

1. **[10]**

(a) Suppose a vector-space consists of three dimensions represented by three keywords, x, y and z. A vector-space query consists of three keywords: x, y, z. In no more than 2-3 sentences, explain the term independence assumption and its drawback.

Term independent assumption states that terms extracted from documents are independent (statistically uncorrelated). Thus, the terms can be used as orthogonal axes of a high-dimensional space. In reality, terms are mostly correlated (i.e., used together), leading to imperfect ranking of pages that contain highly correlated terms

(b) Suppose x, y and z are, respectively, "computer", "science", "department", which are clearly dependent. Explain the effect of the term-independence assumption on document scores computed by the inner product similarity formula.

Below I give a fairly complete explanation, which I would not expect from the students. The key idea is the effect of dependence or correlation between two terms, thus making the match of a redundant term less significant compared to an independent term.

Background: Under term independence, the three terms are treated as independent, and their scores are computed separately and added together (this is how inner product works).

The short answer is that "computer' and "science" are highly dependent or redundant. For a query "computer" "science", under independence assumption, the partial scores generated by "computer" and "science" are simply added. Since the two terms are dependent, matching "computer" infers a degree of matching "science", and vice versa. The summation of the two matches should be lower than if the two matches are independent.

Consider the query "computer" "department", which are less dependent compared to "computer" and "science", and two documents containing "computer" and "science", and "computer" and "department", under term independence, each document matches one out of the two query terms, so they have equal similarity score. However, based on intuition, the second document is more relevant than the first one.

1. **[12]**

**(a) [4]** In the vector space below, and are two document vectors. Suppose terms *x* and *y* represent two topics and the scales represent the intensities of the topics in the documents. Draw and in the vector space below. You can extend the axes if needed.

See blue and green vectors.

**(b) [2]** What are the expected/estimated contents represented by the resulting vectors?

-10

Term *x*

10

20

10

20

30

30

Term *y*

is expected to talk about both topics represented by *x* and *y*. Since only talks about but not , the resulting document talks more about than .

still talks about both topics represented by *x* and *y* but with weakened intensity about *x*.

**(c) [6]** Assuming that the numbers on the x and y axes indicate the weights of the respective terms, compute the cosine similarity and the centroid of and .

1. **[15]** Suppose there are only 5 unique terms (numbered 1 to 5) in the collection, which contains a total of 10 documents. These five term’s term frequencies in a document *D* and their document frequencies are given below:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**(a) [5]** Write down the document vector when tf/tfmax \* idf weighting is used.  
  
*t*1 = 2/5 \* 3.32 = 1.33  
*t*2 = 0/5 \* 2.32 = 0  
*t*3 = 1/5 \* 1.74 = 0.35  
*t*4 = 5/5 \* 2.32 = 2.32  
*t*5 = 2/5 \* 0 = 0  
  
**(b) [5]** Given the query vector, *Q* = 〈 1, 0, 0, 1, 0 〉, compute the cosine similarity values between *Q* and *D*.  
   
inner product = 1.33 + 2.32 = 3.65  
|Q| = sqrt(2) = 1.414  
|D|= sqrt(1.332 + 0.352 + 2.322) = sqrt(7.27) = 2.70  
Cosine(Q,D) = 3.65 / (1.414 \* 2.70) = 0.956  
  
  
**(c) [5]** By comparing the cost of computing the *Q⋅D* inner product, explain why the normalization factor in cosine similarity is expensive to compute.   
  
Even though the vector space has an extremely high dimension, the computation of is very fast since has only a few keywords (i.e., a few non-zero elements). However, to compute the normalization factor, must be computed, and has many non-zero elements (thousands for a typical article) and thus the computation of ||D|| is expense.

[Not required] As an aside, cannot be pre-computed due to the constant changing of the IDF of a term.

1. **[20]** Given the following Web graph, initial PR values (iteration 0) for all pages are 1/4, and damping factor d=0.8:

**(a) [2]** Show the PR formula for computing the PR of each page, e.g., express PR(A) in terms of PR(B), PR(C) and PR(D).

|  |  |
| --- | --- |
| PR(A) | 0.2+ 0.8\*(PR(C)/2) |
| PR(B) | 0.2+ 0.8\*(PR(A)/2) |
| PR(C) | 0.2+ 0.8\*(PR(A)/2 + PR(B) +PR(D)) |
| PR(D) | 0.2+ 0.8\*(PR(C)/2) |

**(b) [8]** Compute the PR values of the pages in the following table. At the end of each iteration, normalize the PR values with L1 norm and show it in the table. Show some intermediate steps in your computation.

Initial value as 1/4

|  |  |  |  |
| --- | --- | --- | --- |
| **Iteration** | **0** | **1** | **2** |
| PageRank(A) | 1/4 | 0.2+ 0.8\*(0.25/2) = 0.3  After norm: 0.3/1.6 = 0.19 | 0.2+ 0.8\*(0.44/2) = 0.38  After norm: 0.38/1.78 = 0.23 |
| PageRank(B) | 1/4 | 0.2+ 0.8\*(0.25/2) = 0.3  After norm: 0.3/1.6 = 0.19 | 0.2+ 0.8\*(0.26/2) = 0.3  After norm: 0.3/1.78 = 0.17 |
| PageRank(C) | 1/4 | 0.2+ 0.8\*(0.25/2 + 0.25 +0.25) = 0.7  After norm: 0.7/1.6 = 0.44 | 0.2+ 0.8\*(0.26/2 + 0.26 +0.26) = 0.72  After norm: 0.72/1.78 = 0.36 |
| PageRank(D) | 1/4 | 0.2+ 0.8\*(0.25/2) = 0.3  After norm: 0.3/1.6 = 0.19 | 0.2+ 0.8\*(0.44/2) = 0.38  After norm: 0.38/1.78 = 0.23 |

**(c) [5]** Would the PR values converge at d=0.8? Comment on the rank sink on the final PR values of the pages. [I expect you to use reasoning instead of computing the PR values in a large number of iterations.]

Yes, it will converge. The A-B-C sub-graph is a "normal" graph without dangling nodes. The rank sink C-D will not cause the infinite growth of PR because of the damping factor.

**(d) [5]** If d=1, would the PR values converge? Comment on the rank sink on the final PR values of the pages. [I expect you to use reasoning instead of computing the PR values in a large number of iterations.]

No, it will not converge. The A-B-C sub-graph is a "normal" graph without dangling node, so it converges even if d=1. However, C-D sub-graph is a rank sink. Without damping, the PR values of C and D will grow indefinitely.

Note: since we normalize the PR values, the maximum value of a PR value is 1. Thus, a PR value may appear to converge to 1, but it in fact does not converge.

\*\* Note, the notion of convergence can be defined differently. Typically, it mean a PR value converges to a single value. However, in the case of rank sink, the PR values of C and D may swap between each other indefinitely and can be consider non-convergence. Alternately, since they only swap between two definite values, we can only consider it as convergence.

For reference only: iteration 0 to iteration 10

|  |
| --- |
| [1] 0.1875 0.1875 0.4375 0.1875  [1] 0.234375 0.171875 0.359375 0.234375  [1] 0.2148438 0.1835938 0.3867188 0.2148438  [1] 0.2216797 0.1787109 0.3779297 0.2216797  [1] 0.2194824 0.1804199 0.3806152 0.2194824  [1] 0.2201538 0.1798706 0.3798218 0.2201538  [1] 0.2199554 0.1800385 0.3800507 0.2199554  [1] 0.2200127 0.1799889 0.3799858 0.2200127  [1] 0.2199965 0.1800032 0.3800039 0.2199965  [1] 0.2200010 0.1799991 0.3799989 0.2200010  [1] 0.2199997 0.1800002 0.3800003 0.2199997 |

1. **[10]** Given the following list of search results, where a tick indicates that the corresponding result is relevant.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank** | **Relevant** | **Score** | **Precision** | **Recall** |
| 1 | √ | 1 | 1.00 | 0.2 |
| 2 |  | 0 | 0.5 | 0.2 |
| 3 | √ | 2 | 0.667 | 0.4 |
| 4 |  | 3 | 0.5 | 0.4 |
| 5 | √ | 0 | 0.6 | 0.6 |
| 6 |  | 0 | 0.5 | 0.6 |
| 7 |  | 0 | 0.429 | 0.6 |
| 8 |  | 0 | 0.375 | 0.6 |
| 9 | √ | 3 | 0.444 | 0.8 |
| 10 |  | 0 | 0.4 | 0.8 |
| … | … | … |  |  |
| 90 | √ | 2 | 0.056 | 1 |
| … | … | … |  |  |
| 100 |  | 0 | 0.05 | 1 |
|  | |  |  |  |

Assuming that there are totally 100 documents and totally 5 of the documents are relevant:

**(a) [3]** Fill in the precision and recall values for the ranks shown in the table.

**(b) [7]** Draw the precision-recall graph in the style of slide 20 of 000-performance.pptx, and interpolate the graph as in slide 21.

0.2

1.0

0.8

0.6

0.4

0.0

0.2

1.0

0.8

0.6

0.4

0.2

1.0

0.8

0.6

0.4

0.0

0.2

1.0

0.8

0.6

0.4

The plotting of the graph is the same as in the slides. Since in this question, not every row is given, filling out of row 89 is important for finding the "turning corners".

By the way, the "Score" column is not useful for this question.

1. [10] Given the document set below, d's are documents, t's are terms and the cells are the weights of the terms in the documents.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **t1** | **t2** | **t3** |
| **d1** | 1 | 1 | 0 |
| **d2** | 1 | 1 | 0 |
| **d3** | 1 | 0 | 1 |
| **d4** | 1 | 1 | 1 |

**(a) [7]** Using term discrimination method and inner product as similarity measure, identify the first and second terms that you would select as index terms for the document.

There are 6 pairs of documents

|  |  |  |  |
| --- | --- | --- | --- |
|  | t1 | t2 | t3 |
| sim(d1,d2) | 1 | 1 | 0 |
| sim(d2,d3) | 1 | 0 | 0 |
| sim(d3,d4) | 1 | 0 | 0 |
| sim(d1,d3) | 1 | 0 | 0 |
| sim(d1,d4) | 1 | 1 | 0 |
| sim(d2,d4) | 1 | 1 | 0 |
| Average | 1 | 0.5 | 0.16 |

So, t3 is selected as the first index term; that leaves t1 and t2.

|  |  |  |
| --- | --- | --- |
|  | t2 | t3 |
| sim(d1,d2) | 2 | 0 |
| sim(d2,d3) | 0 | 0 |
| sim(d3,d4) | 0 | 2 |
| sim(d1,d3) | 0 | 0 |
| sim(d1,d4) | 2 | 0 |
| sim(d2,d4) | 2 | 0 |
| Average | 1 | 0.33 |

So, t3 is selected.

**(b) [3]** Comment on the independence of the terms selected in (a) based on your understanding of the term independence assumption in the vector space model.

The idea of term discrimination value is "to identify the keywords that would separate the documents as far away as possible". In order words, the selected terms are most dissimilar.