COMP 4321 Search Engine for Web and Enterprise Data

Mid-term Examination, Fall 2014

October 28th, 2014

Time Allowed: 1 Hour

Scores:

1	2	3	4	5	6	7	8

1. **[15 points] Circle** True or False for the following statements.

[T / [✓ F]] It is legal for Google to sell brand names (e.g., BMW, Gucci) as keywords in AdWords to advertisers who do not own the brand names

[T / [✓ F]] Recall is typically not used to evaluate web search engines since they all have good performance to satisfy the users

[T / [✓ F]] Term weights must be determined by the tf×idf weighting method

[T / [✓ F]] In the tf×idf weighting method, a term is most important if it appears in every document.

[[✓ T] / F] Inner product favors long documents

[T / [✓ F]] Jaccard similarity is independent to the number of terms in a document

[T / [✓ F]] Cosine similarity measures the cosine of the angle between the document vector and the origin of the vector space

[T / [✓ F]] Documents containing all of the query terms will always be ranked higher than those containing only some of the query terms

[[✓ T] / F] The damping factor (d) in the PageRank algorithm affects the number of iterations required for the PageRank values to stabilize.

[[✓ T] / F] The rank-sink problem can be avoided by setting a small d value in the PageRank formula.

[T / [✓ F]] The PageRank algorithm helps a search engine to identify pages that are highly relevant to the queries.

[T / [✓ F]] The PageRank of a page depends on the query being processed.

[[✓ T] / F] Both WISE and Hypursuit interpret a link as an indicator of content similarity between two pages.

[T / [✓ F]] Hypursuit computes the similarity between the query and the documents using the hyperlinks between documents.

[T / [✓ F]] It is easier to support relevance feedback in the Vector Space model because both queries and documents are vectors of the same format.

2. **[5 points]** In the pure vector-space model, adding one more query term to a query will:

A. Reduce the total number of results returned
B. Increase the total number of results returned
C. Improve the quality of the results returned
D. Improve the quality of the results on the first result page
E. all of the above
none of the above

B

3. **[10 points]** State ONE main difference between how Google, Hypursuit and Wise exploit links between pages in their ranking algorithms. Be precise

HyPursuit use links to infer the similarity between pages that are linked together and use the similarity to group similar pages together.

Wise use links to pass similarity scores from the parent page to the child page and use the scores to rank pages.

Google uses links to infer the authority or quality of a page that are pointed at by the links.

4. **[15 points]**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:

$tf_{D,1} = 2$ $df_1 = 1$ idf=$\log_2(10/1)=3.32$
$tf_{D,2} = 0$ $df_2 = 2$ idf=$\log_2(10/2)=2.32$
$tf_{D,3} = 1$ $df_3 = 3$ idf=$\log_2(10/3)=1.74$
$tf_{D,4} = 5$ $df_4 = 2$ idf=$\log_2(10/2)=2.32$
$tf_{D,5} = 2$ $df_5 = 10$ idf=$\log_2(10/10)=0$

Write down the document vector when tf/tf_{max} * idf weighting is used.

$t_1 = 2/5 * 3.32 = 1.33$
$t_2 = 0/5 * 2.32 = 0$
$t_3 = 1/5 * 1.74 = 0.35$
$t_4 = 5/5 * 2.32 = 2.32$
$t_5 = 2/5 * 0 = 0$

Given the query vector, $Q = \langle 1, 0, 0, 1, 0 \rangle$, compute the cosine similarity values between *Q* and *D*.

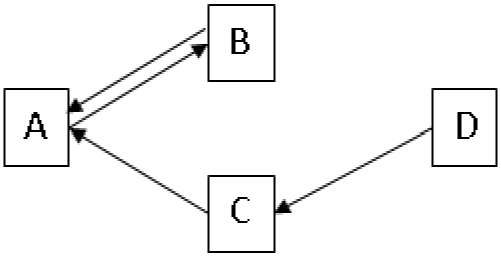
inner product = 1.33 + 2.32 = 3.65
|Q| = sqrt(2) = 1.414

|D|= sqrt(1.33^2 + 0.35^2 + 2.32^2) = sqrt(7.27) = 2.70
Cosine(Q,D) = 3.65 / (1.414 * 2.70) = 0.956

[5] Explain why cosine similarity is expensive to compute.

Because the length of each document vector must be computed in cosine similarity and it cannot be pre-computed due to the constant changing of the IDF of a term.

5. **[35 points]** You are given the web graph below.
(a) Compute the Hub and Authority weights of the pages, normalizing the results in each iteration with the vector length.
(b) Compute the PageRank for d=0.5 For both (a) and (b), compute the values for the first 3 iterations, correct to 2 decimal places, given the initial weights are 1 in iteration 0.



a) [10] Authority Weights:
Authority Weights: [summation of hub weights of parents]

	0		1	Normalized	2	Normalized
A	1	Hub(B)+Hub(C)	2	0.82	2	0.82
B	1	Hub(A)	1	0.41	1	0.41
C	1	Hub(D)	1	0.41	1	0.41
D	1	0	0	0	0	0

Hub Weights: [summation of authority weights of children]

	0		1	Normalized	2	Normalized
A	1	Aut(B)	1	0.5	1	0.32
B	1	Aut(A)	1	0.5	2	0.64
C	1	Aut(A)	1	0.5	2	0.64
D	1	Aut(C)	1	0.5	1	0.32

(b) [10] PageRank:

Iteration	**0**	**1**	**2**

PageRank(A)	1	0.5+0.5*(1/1+1/1)=1.5	0.5+0.5*(1/1+1/1)=1.5
PageRank(B)	1	0.5+0.5*(1/1)=1	0.5+0.5*(1.5/1)=1.25
PageRank(C)	1	0.5+0.5*(1/1)=1	0.5+0.5*(0.5/1)=0.75
PageRank(D)	1	0.5+0.5*0=0.5	0.5

(c) [10] What is a rank sink? Is there a rank sink in the web graph? What is the impact of the rank sink if d<1 and d=1?

A rank sink is a pair of pages linking to each other. There is a rank sink A-B in the given web graph. Rank sink will indefinitely accumulate pagerank. The damping factor of d< 1 will damp the pagerank, but when d=1 there is no damping and the pagerank of the pair will grow indefinitely.

6. **[8 points]** Document D contains the following content: **Hong Kong University** of **Scienc**e and **Technology** is the **best** in the **world**, where significant words are in bold. (i) **[2]** Does D satisfy the "term independence assumption" in vector space model? (ii) **[6]** Explain the impact of term independent assumption on D's ranking for the query: best university in science and technology.

The term independence assumption says that terms in the documents occur independently. The fact that Hong Kong University of Science and Technology is a name, the words occur together most of the time. Thus, D does not satisfy the assumption.

There are 4 significant terms in the query, matching four out of 7 significant terms in D. Common similarity measures would result in high similarity scores between the query and D. However, "university", "science" and "technology" always occur together in D representing one object, but they are counted as three matches. It is not fair.

7. **[5 points]** Give one disadvantage of the design of Clever's architecture. Give a brief justification.

Anyone will do:
- Query time is long because hub and authority weights are computed when the query is being processed
- The subgraph is too small for getting meaningful hub and authority weights
- It relies on a pretty good search engine to retrieve the root set. Since the root set is small, the search engine gets bad results, the subgraph is not a good representation of the query topic.

8. **[7 points]** You have a website consist of two pages A and B that you can freely edit. Suppose A has a link pointing to B. There are many other pages on the web but you cannot modify them. Is it possible to increase the Authority of B? If yes, how; if no, why?

Yes, just create some links in A pointing to other pages on the web. These links will increase the hub weight of A, and hence the authority weight of B.