Iowa State University, Computer Science Department

Computer Science 535
Project 3

Yaojie Jason Hu, Joo Seung Song

November 26, 2019

1 Page rank number of iterations to convergence

Please see table 1.1 for results

$\beta \backslash \epsilon$	0.01	0.0001
0.25	3	5
0.85	7	21

Table 1.1: Iterations taken to converge

2 Trust rank comparison

The trust rank is numerically close to the page rank. The sum of trust rank and page rank are both 1. The absolute difference between the two ranks are 0.34 in our experiment. The trust rank tracks the page rank, mostly.

The top 10 pages in page rank is [354, 3506, 4913, 62, 3798, 4966, 386, 1029, 3516, 27].

The top 10 pages in trust rank is [354, 3506, 62, 4913, 3798, 4966, 31, 3516, 1029, 386].

The order of trust rank and page rank are different. However, given the top 10 pages, it seems that the pages with higher page ranks generally have higher trust ranks.

The bottom 10 pages in page rank is [4837, 4851, 5303, 5368, 5370, 5373, 5415, 5425, 5459, 5465].

The bottom 10 pages in trust rank is [5303, 5368, 5370, 5373, 5415, 5425, 5459, 5465, 5468, 5475].

The pages with lower page ranks also have lower trust ranks.

3 Spam farm

We created three spam farms, with number of spam pages set to be 10, 100, and 1000. The target page is selected to be page 5304, which has the least page rank without spam farm.

The increase of the page rank and trust rank after creation of spam farm is in table 3.1. As the number of spam farms increase, the page rank of the target page increases significantly. As the number of spam farm becomes 100, the page rank of the target page becomes the highest among all pages.

The trust rank changes as spam farm increases, but the correlation is not clear, and in fact, the trust rank decreases as the number of spam farms increase from 10 to 100.

Given the experiment result, we can determine if a page is a spam page if there is a significant difference between the page rank rank and the trust rank rank.

spam farms	page rank	trust rank	highest PR	highest TR	PR rank	TR rank
0	1.511e-5	9.142e-8	4.272e-3	4.261e-3	9965	9963
10	5.889e-4	5.296e-4	4.272e-3	4.261e-3	117	162
100	7.138e-3	1.570e-4	7.138e-3	4.444e-3	1	2281
1000	6.003e-2	2.378e-4	6.003e-2	4.429e-3	1	1063

Table 3.1: Effect of spam farms on page rank and trust rank. Highest page rank or trust rank is the highest value among all pages. The PR rank and TR rank is the ranking of the target page among all pages.

4 Positional Index queries

We tested the following queries from table 4.1 to table 4.5.

We went back to the files and found the phrases to appear in the files as expected. However, TPScore seems to have a strong influence of the Relevance. While pairs of words do appear in the documents, they oftentimes scatter in different locations which do not have any semantic relationships with the queried phrase. This happens especially often when the query contains propositional words, e..g. the query "members of the 500".

Moreover, we see that with short queries, the behavior of TPScore and VSScore both produce no informative results. With one word query, the TPScore is zero, because there are no pairs of words, while VSScore is only binary, which indicate whether the document contains the query word or not.

Despite room for improvements, the ranking produced is correct and acceptable given the algorithm and formulas.

Document	Relevance	TPScore	VSScore
117th_IOC_Session.txt	0.40000	0.00000	1.00000
12U_Baseball_World_Cup.txt	0.40000	0.00000	1.00000
15U_Baseball_World_Cup.txt	0.40000	0.00000	1.00000
16-inch_softball.txt	0.40000	0.00000	1.00000
1845_to_1868_in_baseball.txt	0.40000	0.00000	1.00000
1857_in_baseball.txt	0.40000	0.00000	1.00000
1857_in_sports.txt	0.40000	0.00000	1.00000
1859_in_sports.txt	0.40000	0.00000	1.00000
1860_in_baseball.txt	0.40000	0.00000	1.00000
1860_in_sports.txt	0.40000	0.00000	1.00000

Table 4.1: Query: national

Document	Relevance	TPScore	VSScore
Davey_Johnson.txt	1.60000	2.00000	1.00000
Jack_Morris.txt	1.60000	2.00000	1.00000
Throwback_uniform.txt	1.60000	2.00000	1.00000
Tony_Gwynn.txt	1.60000	2.00000	1.00000
Howard_Cosell.txt	1.60000	2.00000	1.00000
Joe_Girardi.txt	1.60000	2.00000	1.00000
Mickey_Cochrane.txt	1.60000	2.00000	1.00000
Umpire_(baseball).txt	1.60000	2.00000	1.00000
California_Golden_Bears.txt	1.59999	2.00000	0.99997
H&C3&A9ctor_L&C3&B3pez.txt	1.59999	2.00000	0.99997

Table 4.2: Query: world series

Document	Relevance	TPScore	VSScore
Honus_Wagner.txt	1.30000	1.50000	0.99999
Grover_Cleveland_Alexander.txt	1.29998	1.50000	0.99996
William_DeWitt,_Jr..txt	1.29998	1.50000	0.99996
Alvin_Dark.txt	1.29998	1.50000	0.99995
Curse_of_the_Bambino.txt	1.29998	1.50000	0.99995
Jos&C3&A9_Ram&C3&ADrez_(infielder).txt	1.29998	1.50000	0.99995
Juan_Lagares.txt	1.29998	1.50000	0.99995
Paul_Blair_(baseball).txt	1.29998	1.50000	0.99995
Philip_Humber&27s_perfect_game.txt	1.29998	1.50000	0.99995
Rocco_Baldelli.txt	1.29998	1.50000	0.99995

Table 4.3: Query: Major League Baseball

Document	Relevance	TPScore	VSScore
40&E2&80&9340_club.txt	1.10448	1.33333	0.76119
3,000_hit_club.txt	1.10026	1.33333	0.75065
20&E2&80&9320&E2&80&9320_club.txt	1.09911	1.33333	0.74777
Nellie_Fox.txt	0.78978	0.80000	0.77444
Comparison_between_cricket_and_baseball.txt	0.74891	0.80000	0.67227
Comparison_of_baseball_and_cricket.txt	0.74891	0.80000	0.67227
Comparison_of_cricket_and_baseball.txt	0.74891	0.80000	0.67227
History_of_the_Philadelphia_Phillies.txt	0.67807	0.57143	0.83804
Satchel_Paige.txt	0.67442	0.57143	0.82891
Hardy_Richardson.txt	0.66994	0.57143	0.81770

Table 4.4: Query: the group of batters

Document	Relevance	TPScore	VSScore
Cincinnati_Reds.txt	1.15216	1.33333	0.88039
Melbourne.txt	1.14372	1.33333	0.85930
1989_Loma_Prieta_earthquake.txt	1.14326	1.33333	0.85814
Madonna_(entertainer).txt	1.14269	1.33333	0.85673
History_of_New_York.txt	1.14090	1.33333	0.85224
Atlantic_City,_New_Jersey.txt	1.13488	1.33333	0.83720
Detroit_Tigers.txt	1.13447	1.33333	0.83616
Montreal_Expos.txt	1.13373	1.33333	0.83433
Texas_Rangers_(baseball).txt	1.13032	1.33333	0.82580
Washington_Senators_(1961&E2&80&931971).txt	1.13032	1.33333	0.82580

Table 4.5: Query: members of the 500