Web Search Engines — Problem Set 2

Himaja Rachakonda
N14633788
hr970

March 7, 2016

1 Problem 1

Given term-document matrix -

	Doc1	Doc2	Doc3	Doc4
Walrus	10	0	0	10
Carpenter	8	0	40	0
Bread	4	24	0	20
Butter	1	16	0	0

$$w(t,d) = \begin{cases} 1 + \log_2 f(t,d) & iff(t,d) > 0 \\ 0 & iff(t,d) = 0 \end{cases}$$

$$i(t) = \log_2(c/o(t))$$

$$\vec{d} = w(t,d) * i(t)$$

Calculating $f(t,d)$,$w(t,d)$, $\vec{d}$ for each of the terms given -

Walrus $o(t) = 2$, $c = 4$, $i(t) = 1$

	$f(t,d)$	$w(t,d)$	$\vec{d}$
Doc1	10	4.32	4.32
Doc2	0	0	0
Doc3	0	0	0
Doc4	10	4.32	4.32

Carpenter $o(t) = 2$, $c = 4$, $i(t) = 1$

	$f(t,d)$	$w(t,d)$	$\vec{d}$
Doc1	8	4	4
Doc2	0	0	0
Doc3	40	6.32	6.32
Doc4	0	0	0

Bread $o(t) = 3$, $c = 4$, $i(t) = \log_2(4/3)$

	$f(t,d)$	$w(t,d)$	$\vec{d}$
Doc1	4	3	1.233
Doc2	24	5.58	2.29
Doc3	0	0	0
Doc4	20	5.32	2.19

Butter $o(t) = 2$, $c = 4$, $i(t) = 1$

	$f(t,d)$	$w(t,d)$	$\vec{d}$
Doc1	1	0	0
Doc2	16	4	4
Doc3	0	0	0
Doc4	0	0	0

So, the Document vectors with each of these terms as a dimension is as follows -

	Doc1	Doc2	Doc3	Doc4
Walrus	4.32	0	0	4.32
Carpenter	4	0	6.32	0
Bread	1.233	2.29	0	2.19
Butter	0	4	0	0

Normalized document vector is as follows-

	Doc1	Doc2	Doc3	Doc4
Walrus	0.707	0	0	0.707
Carpenter	0.535	0	0.85	0
Bread	0.1066	0.198	0	0.189
Butter	0	1	0	0

1.1 Query - Document Rankings

Query - "Walrus" $\vec{q} =< 1, 0, 0, 0 >$

	$sim(\vec{d}, \vec{q})$	Rank
Doc1	0.707	1
Doc2	0.535	2
Doc3	0.1066	3
Doc4	0	4

Query - "Walrus Carpenter" $\vec{q}$= <0.707,0.707,0,0>

	$sim(\vec{d}, \vec{q})$	Rank
Doc1	0.499	2
Doc2	0.378	3
Doc3	0.2153	4
Doc4	0.707	1

Query - "Walrus Bread Butter" $\vec{q}$= <0.57,0,0.57,0.57>

	$sim(\vec{d}, \vec{q})$	Rank
Doc1	0.806	1
Doc2	0	4
Doc3	0.2813	3
Doc4	0.57	2

2 Problem 2

2.1 Document Similarity

$sim(\vec{d_1}, \vec{d_2})$ = (0.707*0 + 0.535*0 + 0.1066*0.198 + 0*1) = 0.0211
$sim(\vec{d_1}, \vec{d_3})$ = (0.707*0 + 0.535*0.85 + 0.1066*0 + 0*0) = 0.45
$sim(\vec{d_1}, \vec{d_4})$ = (0.707*0.707 + 0.535*0 + 0.1066*0.189 + 0*0) = 0.52

3 Problem 3

4 Problem 4

4.1 $N = 9$, $e = 0.3$, $f = 1 - e \Rightarrow f = 1 - 0.3 \Rightarrow f = 0.7$, $E = (e/N) \Rightarrow E = 0.033$

$$\begin{aligned}
A &= 0.033 + 0.7(0) \\
B &= 0.033 + 0.7(A/4 + C/3) \\
C &= 0.033 + 0.7(A/4 + I/2 + B/2) \\
D &= 0.033 + 0.7(A/4 + H/1) \\
E &= 0.033 + 0.7(A/4 + B/2 + C/3 + F/2 + D/2) \\
F &= 0.033 + 0.7(C/3 + E/2) \\
G &= 0.033 + 0.7(D/2) \\
H &= 0.033 + 0.7(E/2 + G/1 + I/2) \\
I &= 0.033 + 0.7(F/2)
\end{aligned}$$

4.2 Page Rank computation

$$Q = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.175 & 0 & 0.233 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.175 & 0.35 & 0 & 0 & 0 & 0 & 0 & 0 & 0.35 \\
0.175 & 0 & 0 & 0 & 0 & 0 & 0 & 0.7 & 0 \\
0.175 & 0.35 & 0.233 & 0.35 & 0 & 0.35 & 0 & 0 & 0 \\
0 & 0 & 0.233 & 0 & 0.35 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.35 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.35 & 0 & 0.7 & 0 & 0.35 \\
0 & 0 & 0 & 0 & 0 & 0.35 & 0 & 0 & 0
\end{bmatrix}$$

To solve these system of equations, we represent these in the form $\vec{c} = B\vec{p}$ Therefore the system of equations can be represented as:

$$\begin{aligned}
A &= 0.033 \\
-0.175A + B - 0.233C &= 0.033 \\
-0.175A - 0.35B + C - 0.35I &= 0.033 \\
-0.175A + D - 0.7H &= 0.033 \\
-0.175A - 0.35B - 0.233C - 0.35D + E - 0.35F &= 0.033 \\
-0.233C - 0.35E + F &= 0.033 \\
-0.35D + G &= 0.033 \\
-0.35E - 0.7G + H - 0.35I &= 0.033 \\
-0.35F + I &= 0.033
\end{aligned}$$

```
q = [0,0,0,0,0,0,0,0,0;
0.175,0,0.233,0,0,0,0,0,0;
0.175,0.35,0,0,0,0,0,0,0.35;
0.175,0,0,0,0,0,0,0.7,0;
0.175,0.35,0.233,0.35,0,0.35,0,0,0;
0,0,0.233,0,0.35,0,0,0,0;
0,0,0,0.35,0,0,0,0,0;
0,0,0,0,0.35,0,0.7,0,0.35;
0,0,0,0,0,0.35,0,0,0];

b = eye(9) − q;

c = ones(9,1);

c = c * 0.033;

p = b \ c;

p =

    0.0330
    0.0586
    0.0849
    0.1686
    0.1784
    0.1152
    0.0920
    0.1855
    0.0733
```

5 Problem 5

5.1 $N = 9$, $e = 0.99$, $f = 1 - e \Rightarrow f = 1 - 0.99 \Rightarrow f = 0.01$, $E = (e/N) \Rightarrow E = 0.11$

$$\begin{aligned}
A &= 0.11 + 0.01(0) \\
B &= 0.11 + 0.01(A/4 + C/3) \\
C &= 0.11 + 0.01(A/4 + I/2 + B/2) \\
D &= 0.11 + 0.01(A/4 + H/1) \\
E &= 0.11 + 0.01(A/4 + B/2 + C/3 + F/2 + D/2) \\
F &= 0.11 + 0.01(C/3 + E/2) \\
G &= 0.11 + 0.01(D/2) \\
H &= 0.11 + 0.01(E/2 + G/1 + I/2) \\
I &= 0.11 + 0.01(F/2)
\end{aligned}$$

5.2 Page Rank Computaion

$$
Q \quad = \quad \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.0025 & 0 & 0.0033 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.0025 & 0.005 & 0 & 0 & 0 & 0 & 0 & 0 & 0.005 \\
0.0025 & 0 & 0 & 0 & 0 & 0 & 0 & 0.01 & 0 \\
0.0025 & 0.005 & 0.0033 & 0.005 & 0 & 0.005 & 0 & 0 & 0 \\
0 & 0 & 0.0033 & 0 & 0.005 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.005 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.005 & 0 & 0.01 & 0 & 0.005 \\
0 & 0 & 0 & 0 & 0 & 0.005 & 0 & 0 & 0
\end{bmatrix}
$$

To solve these system of equations, we represent these in the form $\vec{c} = B\vec{p}$ Therefore the system of equations can be represented as:

$$
\begin{aligned}
A &= 0.11 \\
-0.0025A + B - 0.0033C &= 0.11 \\
-0.0025A - 0.005B + C - 0.005I &= 0.11 \\
-0.0025A + D - 0.01H &= 0.11 \\
-0.0025A - 0.005B - 0.0033C - 0.005D + E - 0.005F &= 0.11 \\
-0.0033C - 0.005E + F &= 0.11 \\
-0.005D + G &= 0.11 \\
-0.005E - 0.01G + H - 0.005I &= 0.11 \\
-0.005F + I &= 0.11
\end{aligned}
$$

```
q = [0,0,0,0,0,0,0,0,0;
0.0025,0,0.0033,0,0,0,0,0,0;
0.0025,0.005,0,0,0,0,0,0,0.005;
0.0025,0,0,0,0,0,0,0.01,0;
0.0025,0.005,0.0033,0.005,0,0.005,0,0,0;
0,0,0.0033,0,0.005,0,0,0,0;
0,0,0,0.005,0,0,0,0,0;
0,0,0,0,0.005,0,0.01,0,0.005;
0,0,0,0,0,0.005,0,0,0];

b = eye(9) - q;

c = ones(9,1);

c = c * 0.11;

p = b \ c

p =

    0.1100
    0.1106
```

```
0.1114
0.1114
0.1123
0.1109
0.1106
0.1122
0.1106
```

5.3 $N = 9$, $e = 0.01$, $f = 1 - e \Rightarrow f = 1 - 0.01 \Rightarrow f = 0.99$, $E = (e/N) \Rightarrow E = 0.001$

$$
\begin{aligned}
A &= 0.001 + 0.99(0)\\
B &= 0.001 + 0.99(A/4 + C/3)\\
C &= 0.001 + 0.99(A/4 + I/2 + B/2)\\
D &= 0.001 + 0.99(A/4 + H/1)\\
E &= 0.001 + 0.99(A/4 + B/2 + C/3 + F/2 + D/2)\\
F &= 0.001 + 0.99(C/3 + E/2)\\
G &= 0.001 + 0.99(D/2)\\
H &= 0.001 + 0.99(E/2 + G/1 + I/2)\\
I &= 0.001 + 0.99(F/2)
\end{aligned}
$$

5.4 Page Rank Computation

$$
Q = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0.2475 & 0 & 0.33 & 0 & 0 & 0 & 0 & 0 & 0\\
0.2475 & 0.495 & 0 & 0 & 0 & 0 & 0 & 0 & 0.495\\
0.2475 & 0 & 0 & 0 & 0 & 0 & 0 & 0.99 & 0\\
0.2475 & 0.495 & 0.33 & 0.495 & 0 & 0.495 & 0 & 0 & 0\\
0 & 0 & 0.33 & 0 & 0.495 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0.495 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0.495 & 0 & 0.99 & 0 & 0.495\\
0 & 0 & 0 & 0 & 0 & 0.495 & 0 & 0 & 0
\end{bmatrix}
$$

To solve these system of equations, we represent these in the form $\vec{c} = B\vec{p}$ Therefore the system of equations can be represented as:

$$
\begin{aligned}
A &= 0.001\\
-0.2475A + B - 0.33C &= 0.001\\
-0.2475A - 0.495B + C - 0.495I &= 0.001\\
-0.2475A + D - 0.99H &= 0.001\\
-0.2475A - 0.495B - 0.33C - 0.495D + E - 0.495F &= 0.001\\
-0.33C - 0.495E + F &= 0.001\\
-0.495D + G &= 0.001\\
-0.495E - 0.99G + H - 0.495I &= 0.001\\
-0.495F + I &= 0.001
\end{aligned}
$$

```
q = [0,0,0,0,0,0,0,0,0;
0.2475,0,0.33,0,0,0,0,0,0;
0.2475,0.495,0,0,0,0,0,0,0.495;
0.2475,0,0,0,0,0,0,0.99,0;
0.2475,0.495,0.33,0.495,0,0.495,0,0,0;
0,0,0.33,0,0.495,0,0,0,0;
0,0,0,0.495,0,0,0,0,0;
0,0,0,0,0.495,0,0.99,0,0.495;
0,0,0,0,0,0.495,0,0,0];

b = eye(9) - q;

c = ones(9,1);

c = c * 0.001;

p = b \ c

p =

    1.0000
   11.4695
   30.9758
  215.7781
  171.5447
   96.1366
  107.8101
  216.6975
   48.5876
```

Processing the posting lists in the order of size is **NOT** optimal in a scenario where the intermediate posting list is either very small or infact zero. This might be the case when there are, for instance 4 posting lists namely A,B,C, and D with posting list sizes in the order of 10, 100, 1M, 10M. In this scenario, if $A \cap D = 0$. Here in this case the complexity will be $O(10 + 100 + 1M + 10M)$ however there would have been no necessity to traverse the list if this was implemented differntly.

5.5 The time given in MRS for the procedure INTERSECT(p1,p2) in figure 1.6 is x+y where x is the length of p1 and y is length of p2. If you use additionally a hash table in which the key is the pair $<$ word, docID $>$, and you record with each word the length of its posting list, then this can be made significantly faster. How do you use the hash table, and how fast does the revised algorithm run?