Link Analysis

Kazi Abir Adnan (2014280162)

Wednesday 29th April, 2015

Introduction

Site-level PageRank Calculation on a given real Web sites sample collection. There are 500,000 sites, 72,584,579 links (1G original / 250M zip file)

Efficiency

Running Time Analysis

- Data Loading Time : 1.66 minutes.
- Time for 1 loop (avg.) : 57.34 seconds.
- Time for 15 loops : 15 minutes.
- Total running time till convergence(56 loops) : 53 minutes.

Hardware Configuration

- Platform: ASUS NOTEBOOK
- Windows Edition: Windows 8.1 Pro
- Processor: Intel(R) Core(Tm) i5-4200U CPU @ 1.60 GHz 2.30 GHz
- Installed Memory (RAM): 8.00 (7.89 GB Usable)
- System Type: 64-bit Operating System, x64-based Processor
- Number of CPUs: 4
- Number of Cores: 2
- L1 Cache size: 128 KB
- L2 Cache size: 512 KB
- L3 Cache size: 3MB

Platform

- JAVA
- JDK 1.8

Analysis of Top ranked 5 sites

My top five ranked sites are -

id	url	PageRank
11	http://www.facebook.com/	0.006518429
3	http://www.miibeian.gov.cn/	0.006360299
13	http://twitter.com/	0.005855289
15	http://www.adobe.com/	0.003896635
64	http://www.linkbucks.com/	0.0029227857

Rank 1

- ID : 11
- URL: http://www.facebook.com/
- Facebook is the most popular social network. It has billions of user. People, Organizations, Celebrities share their profile and it increases in link count of this website.

Rank 2

- ID : 3
- URL: http://www.miibeian.gov.cn/
- It is a government website. So, people may use this site as reference in other websites and increases the link count.

Rank 3

- ID : 13
- URL: http://twitter.com/
- It is a social network and very much popular. People share tweets and the links. So, it should have Page Rank high.

Rank 4

- ID : 15
- URL: http://www.adobe.com/
- Websites which uses flash might link to this website. So, surfers click this link to download flash player. So, it is logical to have Page Rank high for this site.

Rank 5

- ID : 64
- URL: http://www.linkbucks.com/
- It is an advertising website. So, People may click the advertisements from other websites intentionally or unintentionally to this link and in link count increases for this website. So, Page Rank that is why is high.

Comparison between loop 5 and Loop 15

Comparison between two situation

Rank	After 5 loops		After 15 loops	
	ID	PageRank	ID	PageRank
1	3	0.0066475477	11	0.0065022544
2	11	0.006153143	3	0.0063805603
3	13	0.0054975324	13	0.005839587
4	64	0.0041185906	15	0.003894141
5	15	0.0037744546	64	0.0030636552
6	33	0.002638819	33	0.002809418
7	80	0.0023803988	80	0.0019139511
8	14	0.0020320166	14	0.0018801702
9	47	0.0017898714	47	0.001871376
10	2	0.0016243939	55	0.0015948894
11	21	0.0016097663	21	0.001519564
12	55	0.0015531913	2	0.001505827
13	0	0.0014023043	0	0.0013098766
14	28	0.0013390576	28	0.0012706841
15	5	0.0012105294	5	0.001132658
16	4	0.0011414869	85	0.0011011268
17	85	0.0010594508	4	0.0010611063
18	36	9.4424334E-4	125	9.487728E-4
19	88	9.04255E-4	112	9.425815E-4
20	112	8.970836E-4	115	8.8914105E-4

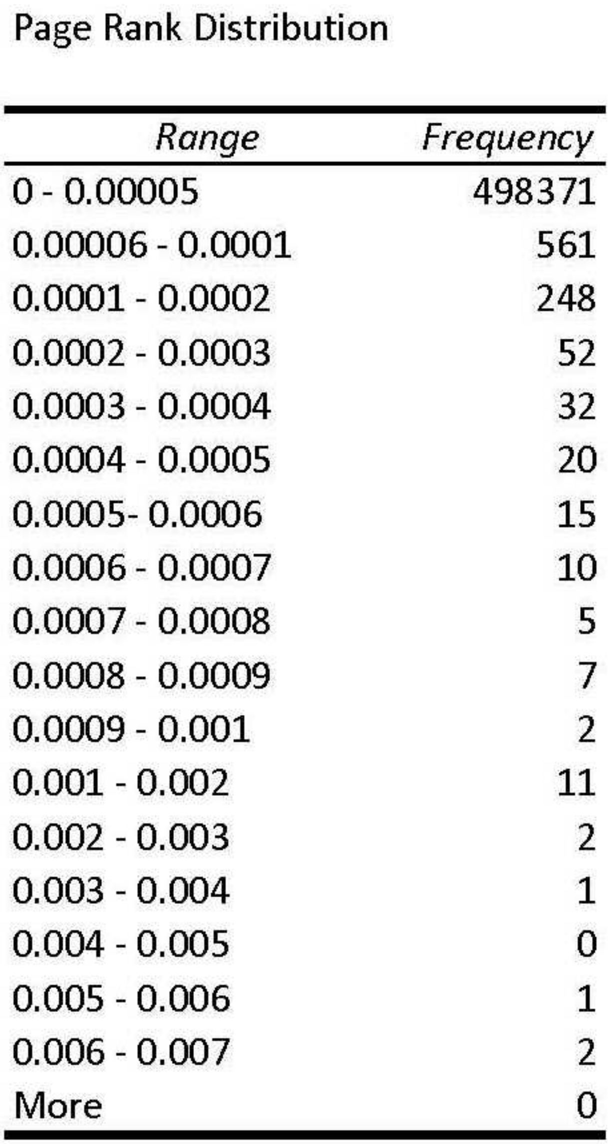
Observation

- I was wondering a scenario. Suppose a very popular site like Facebook or Google. People who uses Internet knows the link and surfs the site directly. How this hit count contributes to measure PageRank of a link? As, in our algorithm I think this scenario is overlooked and it doesn't contribute to the calculation.
- Page Rank doesn't help to recognize spam pages.
- The value of Page Rank is very small. And if we see the distribution of the page ranks, most of the page ranks are very small to be used. Only few of the sites have smart page ranks and other values are really unnecessay.

Discussion

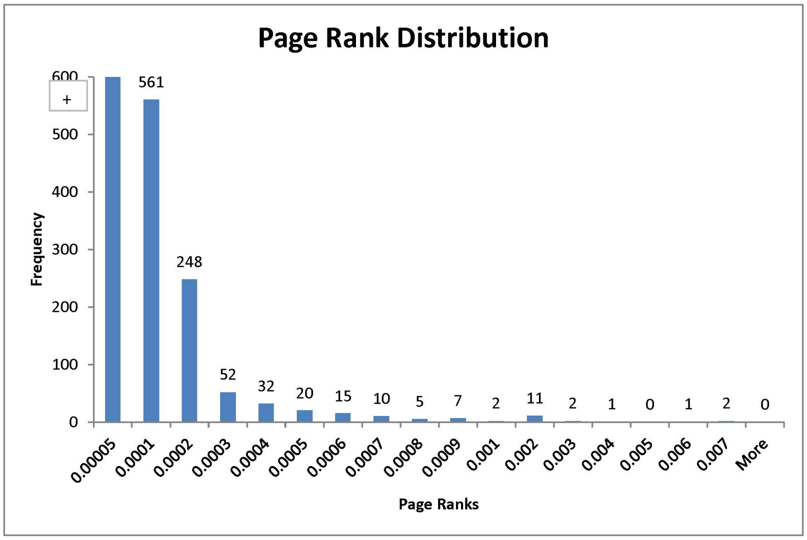
- I have used ArrayList instead of HashMap Or Set. ArrayList is much faster than mapping here. Because, we are accessing the data linearly here.
- As, id's are not consecutive here, I have mapped original ID's to consecutive ID's. The advantage is that it helps spatial locality and increases cache hit.
- I had to increase the heap size of the memory to allocate data.
- I have optimized the algorithm by calculating page ranks of no out links once in loop. It reduces the running time in a great manner.

Distribution of PageRank scores

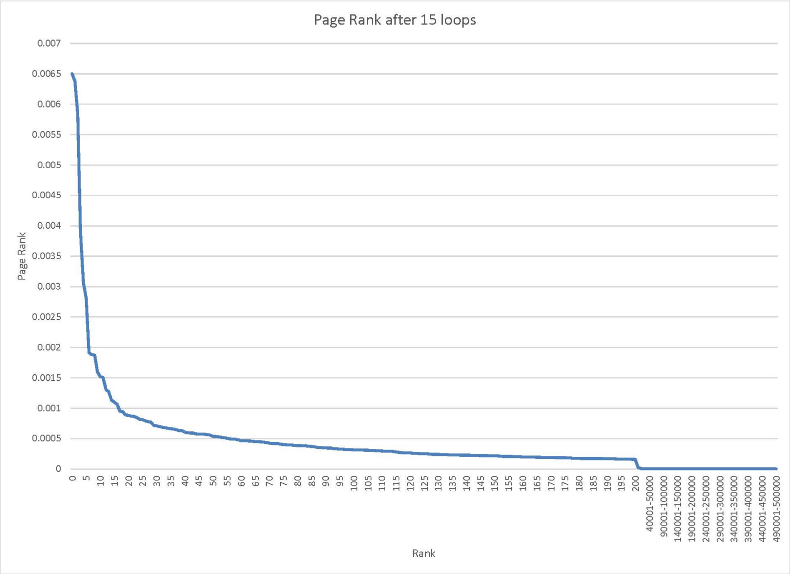


Page Rank Distribution

Range	*Frequency*
0 - 0.00005	498371
0.00006 - 0.0001	561
0.0001 - 0.0002	248
0.0002 - 0.0003	52
0.0003 - 0.0004	32
0.0004 - 0.0005	20
0.0005- 0.0006	15
0.0006 - 0.0007	10
0.0007 - 0.0008	5
0.0008 - 0.0009	7
0.0009 - 0.001	2
0.001 - 0.002	11
0.002 - 0.003	2
0.003 - 0.004	1
0.004 - 0.005	0
0.005 - 0.006	1
0.006 - 0.007	2
More	0



Page Rank Distribution
Frequency
600
500
400
300
200
100
0
561
248
52
32
20
15
10
5
7
2
11
2
1
0
1
2
0
0.00005
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
0.0007
0.0008
0.0009
0.001
0.002
0.003
0.004
0.005
0.006
0.007
More
Page Ranks



Page Rank after 15 loops
Page Rank
0.007
0.0065
0.006
0.0055
0.005
0.0045
0.004
0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
160
165
170
175
180
185
190
195
200
40001-500000
90001-1000000
1400001-1500000
1900001-2000000
2400001-2500000
2900001-3000000
3400001-3500000
3900001-4000000
4400001-4500000
4900001-5000000
Rank