Text Based Information Retrieval (H02C8A) Link-based algorithms

Li Quan

March 31, 2011

1 Calculating PageRank

We calculate the PageRank vector for the graph of *sample-tiny.txt*, visualised in Figure 1). First we use no damping factor—i.e., with the basic iterative power iteration, where we completely preserve the given web structure.

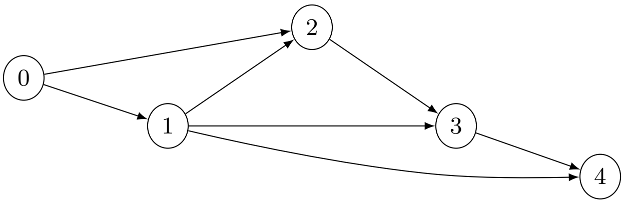

Figure 1: The web graph of *sample-tiny.txt*.

Table 1 shows the PageRank values after each iteration: after 4 iterations, the result is the null vector. This result is obviously meaningless. Node 4 drains all the importance from the web, in each iterative step taking some of the importance of nodes 1 and 2, but not passing it on to other nodes, as it has no outgoing links[1].

To fix this, pages with no outbound links are assumed to link out to all other pages in the graph. Now, the hyperlink matrix of the graph is stochastic and, by the Perron–Frobenius theorem, it has a unique stationary probability vector, i.e., its PageRank, which is found using the algorithm after about 15 iterations: $(0.0769, 0.1154, 0.1539, 0.2692, 0.3846)$.

However in general, we also have to use a damping factor α to ensure convergence of the power iteration algorithm (see [Aus11]).

We now calculate the PageRank values of the web graph given in *sample-large2.txt*, with various damping values from 0 to 1 in increments of 0.05. Figure 2 shows the

[1] A node with no outgoing links is a so called dangling node [Aus11].

	PageRank value of node				
i	0	1	2	3	4
0	0.2000	0.2000	0.2000	0.2000	0.2000
1	0.0000	0.1000	0.1667	0.2667	0.2667
2	0.0000	0.0000	0.0333	0.2000	0.3000
3	0.0000	0.0000	0.0000	0.0000	0.3000
4	0.0000	0.0000	0.0000	0.0000	0.0000
⋮	⋮	⋮	⋮	⋮	⋮

Table 1: Distribution of PageRank values of the nodes of the graph *sample-tiny.txt* in the ith iteration.

standard deviation of the PageRank values and Figure 3 the number of iterations for convergence.

We clearly see that for small values of α, the algorithm takes less iterations to converge and the standard deviation is smaller, compared to larger values of α.

Assume a random surfer. With probability α, he follows a link of the current page, with probability $1 - \alpha$ he surfs to a random page. As $\alpha \to 0$, the web has a link between any two pages and we have lost the original hyperlink structure of the web. When $\alpha \to 1$, the original hyperlink structure is taken more in account.

Clearly, important pages (i.e., pages that happen to be linked by many other pages, or by few important ones) will be visited more often when α is large [Aus11]. However, when we take a large value of α, the convergence of the power method will be very slow. Additionally, contrary to popular belief, $\alpha \approx 1$ does not deliver “better” PageRank values [BSV05]. Intuitively this is easily explained: we don’t surf on the web just clicking on hyperlinks all the time.

In conclusion, a good trade-off between meaningful PageRank values and convergence rate is achieved using the frequently used $\alpha = 0.85$.

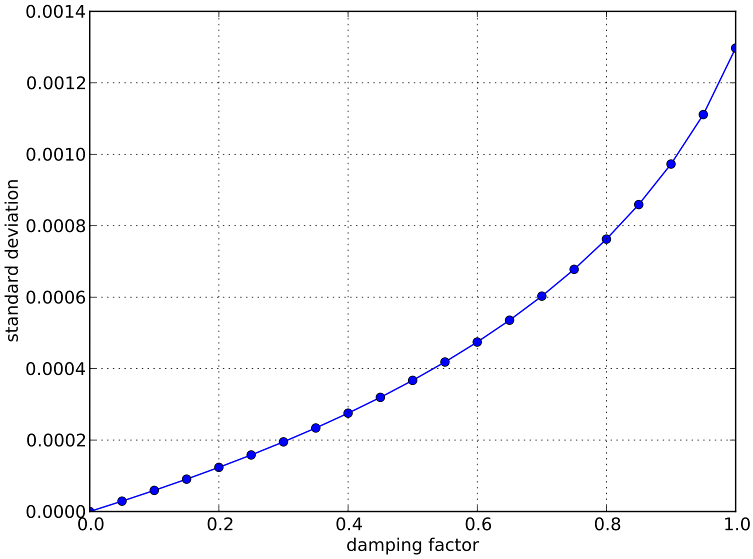

Figure 2: Damping value and standard deviation of PageRank scores.

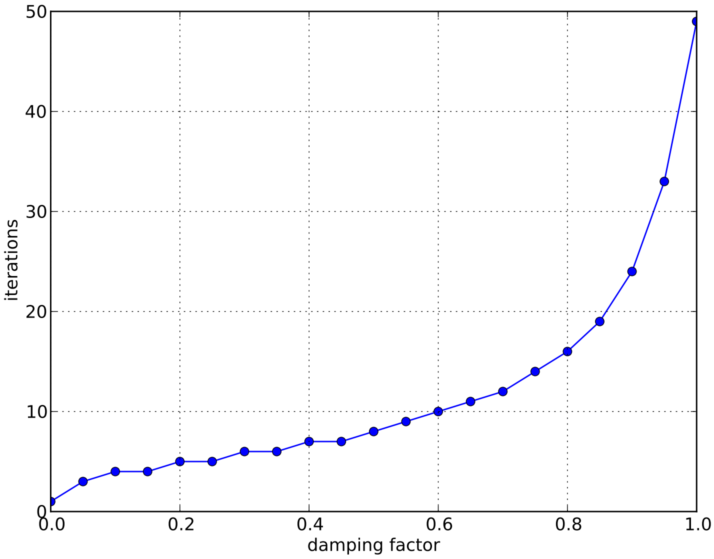

Figure 3: Damping value and number of iterations required for convergence of PageRank scores ($\epsilon = 1 \times 10^{-4}$).

2 Topical Pagerank

Using an initial PageRank vector where the initial values are 0.02 for normal nodes, and 0.5 for the topic X nodes, after

we get following results pagerank of 8614504 is 0.00310119751126 pagerank of 10936880 is 0.0023332770172 pagerank of 8848271 is 0.00196661281637

References

[Aus11] David Austin. How google finds your needle in the web's haystack, 2011. `http://www.ams.org/samplings/feature-column/fcarc-pagerank`.

[BSV05] Paolo Boldi, Massimo Santini, and Sebastiano Vigna. Pagerank as a function of the damping factor. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 557–566, New York, NY, USA, 2005. ACM.

[PBMW98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998. 17 p.