BackRank: an Alternative for PageRank?

Mohamed Bouklit, Fabien Mathieu

BackRank: an Alternative for PageRank?

Mohamed Bouklit
LIRMM
34392 Montpellier Cedex 5 France
bouklit@lirmm.fr

Fabien Mathieu
Gyroweb – INRIA, LIRMM
34392 Montpellier Cedex 5 France
fmathieu@clipper.ens.fr

ABSTRACT

This paper proposes to extend a previous work, *The Effect of the Back Button in a Random Walk: Application for PageRank* [5]. We introduce an enhanced version of the PageRank algorithm using a realistic model for the *Back* button, thus improving the random surfer model. We show that in the special case where the history is bound to an unique page (you cannot use the *Back* button twice in a row), we can produce an algorithm that does not need much more resources than a standard PageRank. This algorithm, BackRank, can converge up to 30% faster than a standard PageRank and suppress most of the drawbacks induced by the existence of pages without links.

Categories and Subject Descriptors: F.2.1 [Analysis of Algorithms and Problem Complexity]: Numerical Algorithms and Problems — *Computations on matrices*

General Terms: Algorithms, Measurement

Keywords: Web analysis, PageRank, Random walk, flow, back button

1. INTRODUCTION

Since the introduction of the *PageRank* ranking in 1998 [6, 2], numerous enhancements were made in both implementation and theorical efficiency of this algorithm [3, 4]. Our purpose is to use a modeling of Back navigation to produce an enhanced PageRank algorithm.

1.1 Notations — Standard PageRank

Let $G = (V, E)$ be a web graph, that is a set V of web pages linked to each other by a set E of hyperlinks.

If G is aperiodic and strongly connected, it is well known that the iterative process

$$\forall v \in V, n \in \mathbb{N}, P_{n+1}(v) = \sum_{w \to v} \frac{P_n(w)}{deg(w)}, \qquad (1)$$

where $deg(v)$ is the out-degree of $v \in V$, converges towards an unique probability P for any initial probability P_0.

However the web graph is far from being strongly connected. One solution is to introduce a damping factor d. The principle is to damp the PageRank flow at each iteration and to redistribute the lost flow according to a given probability Z on V[1]:

$$\forall v \in V, n \in \mathbb{N}, P_{n+1}(v) = d \sum_{w \to v} \frac{P_n(w)}{deg(w)} + \mu_n Z(v), \qquad (2)$$

where $d \in [0, 1]$ is the damping factor and $\mu_n \in [(1-d), 1]$ is such that $\sum_{v \in V} P_{n+1}(v) = 1$.

Using a damping factor is equivalent to working on a weighted strongly connected graph. Note that without dangling nodes, $\mu_n = 1 - d$. Otherwise, μ_n vary, but converges towards an asymptotic value (and so does P_n).

2. BACK BUTTON MODELLING

The approach we follow is similar to [7] [2]. We suggest to refine the PageRank model by incorporating the possibility to *return* with a bounded history stack. Potentially, adding a stochastic process with finite memory m to a Markov chain without memory can lead to consider all the possible paths in G of length m in G.

We will focus on the case $m = 1$. The canonical workspace is then the set E of the hyperlinks. We have introduced two intuitive *Back* button models in [5] for $m = 1$, one of them collapsing the working space from E to V.

Irreversible Back. We suppose that the *Back* button cannot be used twice consecutively: this button is deactivated after its use and it is necessary to use at least once a real link before being able to use it again. This model, which seems more complex, has however important advantages compared to the reversible model[5].

1. It may be more realistic: in real browsers, the Back button is deactivated when the history is empty. Yet, the usage of the Forward button is rather anecdotal, thus leading us to an irreversible model.

2. We found out that the damping factor was easier to introduce with an irreversible Back modeling.

3. The resulting algorithm, BackRank, is almost as easy to implement as a standard PageRank algorithm, and as we will see in Section 3, it converges fast.

4. The Back navigation induces a sort of *greenhouse effect* at the dangling nodes level that is similar to the rank sink phenomenon described in [6]. The irreversible model reduces this effect.

[1] Z is a probability *by default*, also known as *zap* distribution. Most of the time, Z is the uniform distribution, but some have suggested that it would be better to "personalize" it.

[2] In fact, both posters was done independently and each one discovered the work of the other during the thirteenth *WWW* conference.

WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

3. RESULTS

We propose to validate our algorithm by confronting it to the standard PageRank algorithm defined by (2). As proposed in [6], and to cope with BackRank, dangling nodes are suppressed during the main loop, and added back for the last few iterations. Both BackRank and standard PageRank use an uniform distribution on R and $d = 0.85$ by default. The convergence criterion was set to 10^{-10}. We have tested both algorithms using the WebGraph framework [1]. We used 118, 18.5 and 41.3 million page sample of real Web graph from 2001, 2002 and 2004, that will be called WG01, WG02 and WG04[3]. The machine used for testing was a SunFire 880 with Solaris.

Convergence. An important criterion to evaluate a PageRank-like algorithm is to observe how fast it converges. Figure 1 confronts, on a semi-logarithmic scale, the value of the convergence parameter δ after n iterations. The sample used was WG01, but other samples behave the same way (the variation of the number of iterations needed to reach the convergence criterion is less than 2). The surprising result is that BackRank needs only 90 iterations versus 123 for PageRank. This difference, that brings a valuable interest to BackRank, can be explained by Gauss-Seidel optimization embedded in BackRank [5].

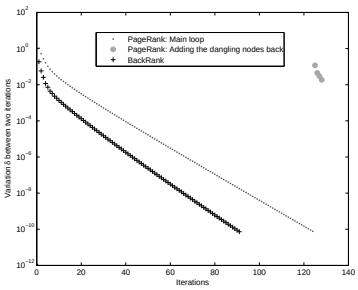

Figure 1: Compared convergences of BackRank and PageRank

Ranking. Technical performances are not the only criteria to evaluate a PageRank-like algorithm. The pertinence of the ranking obtained is also primordial. A first approach is to compare quantitatively the top-ranked pages according to BackRank and PageRank. Figure 2 shows such a comparison: for a given integer n, the overlap percentage between the n top-ranked BackRank pages and the n top-ranked PageRank pages is plotted. If we consider the first 1% top-ranked pages, the overlap varies b etween 90% and 92% according to the studied sample.

To conclude this overview of the quality of the BackRank ranking, Table 1 shows the 20 top-ranked URLs returned by BackRank and PageRank for WG01. For instance, we can notice that `http://www.altavista.com/` appears only in BackRank top 20 (PageRank ranked it #42). On the other hand, `http://www.w3.org/` is missing in BackRank (in fact, its rank is #21). Of course, we cannot affirm that BackRank is better than PageRank until it has been incorporated in an experimental search engine for full scale testings, but Table 1 makes us hope it may be a good challenger.

[3]WG01 has been obtained by the WebBase Project crawler: `http://www-diglib.stanford.edu/~testbed/doc2/WebBase/`
WG02 and WG04 come from the UbiCrawler Project: `http://ubi.imc.pi.cnr.it/projects/ubicrawler/`

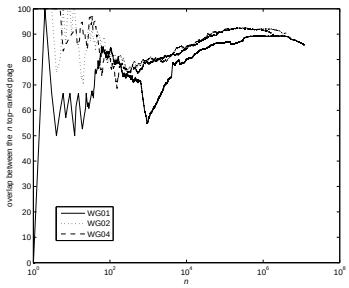

Figure 2: Measure of the overlap between the n top-ranked pages of BackRank and PageRank

Rank	BackRank	PageRank
1	http://www.yahoo.com/	http://www.adobe.com/prodindex/acrobat/readstep.html
2	http://www.adobe.com/prodindex/acrobat/readstep.html	http://www.yahoo.com/
3	http://news.tucows.com/	http://www.worldwidemart.com/scripts/
4	http://www.altavista.com/	http://www.adobe.com/products/acrobat/readstep.html
5	http://www.adobe.com/products/acrobat/readstep.html	http://www.ibm.com/
6	http://home.netscape.com/	http://home.netscape.com/
7	http://www.domaindirect.com/	http://www.listbot.com/
8	http://www.worldwidemart.com/scripts/	http://www.acme.com/software/thttpd/
9	http://www.ibm.com/	http://www.adobe.com/
10	http://www.btsw.com/	http://www.w3.org/
11	http://webcrossing.com/	http://www.adobe.com/homepage.html
12	http://www.real.com/	http://www.adobe.com/misc/privacy.html
13	http://www.acme.com/software/thttpd/	http://www.domaindirect.com/
14	http://www.listbot.com/	http://www.adobe.com/misc/copyright.html
15	http://www.adobe.com/	http://www.adobe.com/misc/comments.html
16	http://www.microsoft.com/windows/ie/default.htm	http://www.adobe.com/store/main.html
17	http://www.macromedia.com/shockwave/download/	http://www.listbot.com/faq.shtml
18	http://counter.rambler.ru/top100/	http://cbl.leeds.ac.uk/nikos/personal.html
19	http://www.mkstats.com/	http://www.listbot.com/cgi-bin/customer
20	http://www.tucows.com/privacy.html	http://news.tucows.com/

Table 1: 20 top-ranked URLs returned by BR and PR on WG01

4. CONCLUSION

BackRank: an alternative for PageRank? For the moment, none of the numerous PageRank optimizations techniques available [3, 4] has been implemented on BackRank. Confronting optimized versions of BackRank and optimized versions of PageRank is the next step in our technical evaluation. The other step in the validation of our BackRank algorithm is to evaluate its quality by incorporating it to a search engine. We have introduced an alternative to standard PageRank algorithms, with a view to use more realistic user patterns in the random surfer modeling. The resulting algorithm, BackRank, behaves better than expected and offers promising perspectives.

5. REFERENCES

[1] P. Boldi and S. Vigna. Webgraph project.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[3] T. Haveliwala. Efficient computation of PageRank. Technical report, Computer Science Department, Stanford University, 1999.

[4] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Extrapolation methods for accelerating pagerank computations. In *Proceedings of the Twelfth International World Wide Web Conference*, 2003.

[5] F. Mathieu and M. Bouklit. The effect of the back button in a random walk: application for pagerank. In *Alternate track papers & posters of the 13th international conference on World Wide Web*, pages 370–371. ACM Press, 2004.

[6] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Computer Science Department, Stanford University, 1998.

[7] M. Sydow. Random surfer with back step. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 352–353. ACM Press, 2004.