Freshness matters: In Flowers, Food and Web Authority

Andreas Sander, Razvan Belet

November 28, 2010

1 Introduction

Freshness in web authority is nowadays a hot topic for search engine providing companies as Google or Yahoo. With the help of thousand new Web 2.0 tools as Wordpress, Wikipedia etc. it was never been easier to create or update content within of seconds.

But search engines often run on a single web snapshot and hence are not able to track temporal changes on a website. Furthermore the ranking algorithms on search engines, as Google, only refer to the number of in-links a website has without considering how fresh, means how new created or updated, the content is. This leads to the fact that older websites, which have more time to gather in-links, rank higher than newer (but fresher) websites with fewer in-links.

The approach presented in the underlying paper [1] wants to fill in this gaps by providing a temporal ranking model, called T-Fresh. To compute an authority score which incorporates freshness, the approach in the paper uses two temporal aspects:

1. web freshness
2. multiple web snapshots at different time points

The following article is divided into two parts. In the first part the approach of the paper is summarised and in the second, strengths, weaknesses and further extensions of this approach are discussed.

2 Summary of the Approach

As described before the aim of Na Dai's and Brian Davison's approach is to deliver a temporal ranking model at the end where web freshness at different time points has an influence on ranking. To achieve this goal, the approach uses several snapshots of the web at different time points.
Furthermore the approach differentiates between the so called page freshness

(abbrev. PF) and in-link freshness (abbrev. InF). For the computation of page or in-link freshness at a time point t_i the corresponding freshness in time point t_{i-1} also has an influence which decays over time. Additionally the freshness score achieved in the interval $[t_{i-1}, t_i]$ must be computed. This is done by considering the corresponding pages in the snapshots at time point t_i and t_{i-1}. Then through a comparison of these two snapshots, the activities performed within this interval must be inferred. In this approach four different link and three page activity types are distinguished (see Table 1), where every action makes a specific, previously defined contribution to the corresponding freshness score (see last column in Table 1).

	Link activity	**Infl. on p's InF**	**Gain of p's InF**
1	creation of link $l : q \rightarrow p$	↑↑↑	3
2	update on link $l : q \rightarrow p$ (changed anchor)	↑↑	2
3	update on link $l : q \rightarrow p$ (unchanged anchor)	↑	1.5
4	removal of link $l : q \rightarrow p$	↓↓	-0.5
	Page activity	**Infl. on q's PF**	**Gain of q's PF**
1	creation of page q	↑↑↑	3
2	update on page q	↑	1.5
3	removal of page q	↓↓	-0.5

Table 1: Activities on pages and links and their influence on web freshness. (The link l points from page q to page p. ↑: positive influence on web freshness. ↓: negative influence on web freshness. The number of ↑ or ↓ indicates the magnitude.)

The in-link/page freshness score achieved in time interval $[t_{i-1}, t_i]$ is further a parameterised combination of two parts:

The first part is an initial in-link/page freshness score achieved by actions on the in-links/pages themselves.
In the second part a specific proportion of in-link/page freshness is inherited by the parent/out-linked pages.

Finally the in-link/page freshness score at time point t_i is then computed as combination of the in-link/page freshness score one time point before and the in-link/page freshness score achieved in the interval $[t_{i-1}, t_i]$.

With the aim to create a time-dependent authority score for every page, the approach uses authority propagation in an archival link graph (see figure 1).

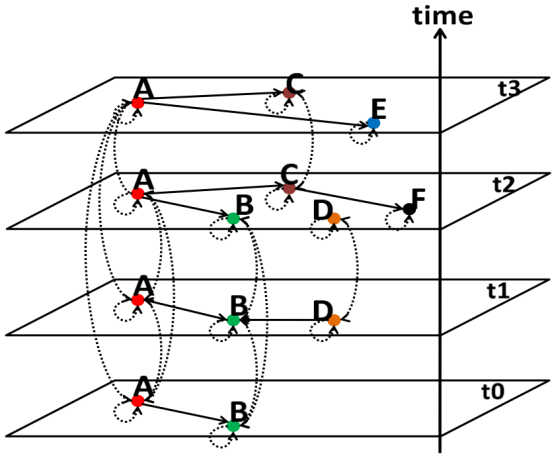

Figure 1: Link graph with four web snapshots at different time points.

This authority propagation is build upon a temporal surfer model, which has a temporal intent and prefers fresh content. In each move at the first step, this surfer model can either follow an out-link of the current page or jump to random page in the same web snapshot, means at the same time point. In a second step the surfer model can choose (based on the temporal intent) a specific time point and the corresponding snapshot of the previously entered page.

The further approach of the paper is based on the fact that a web surfers normal behaviour, by entering pages and browsing this page a specific amount of time, is comparable to a semi-markov process which also enters different states and stays there for a random amount of time.

As a measure for the final authority score the approach adopts a formula which is actually used to compute a long-run proportion of time a semi-markov process is in state i,

$$A(i) = \frac{\pi_i \mu_i}{\sum_{j=1}^{n} \pi_j \mu_j}, i = 1, 2, \cdots, n \tag{1}$$

where n is the total number of states, π_i the probability that the process enters state i and μ_i denotes the mean time the semi-markov process is staying in state i.

Hence for the computation of the final authority score, the authors define two formulas for $\pi_{p,i}$ and $\mu_{p,i}$ (depending on a page p and a time point i), where $\pi_{p,i}$ denotes the probability that a web surfer reaches page p at time point i and $\mu_{p,i}$ representing the mean staying time of a web surfer on page p at snapshot of time point t_i. Thereby page freshness is used to model the web surfers preference for following out-linked pages, which has influence on $\pi_{p,i}$. Moreover in-link freshness is used to compute the staying time $\mu_{p,i}$.

To achieve a formula which computes an authority score for a specific page p at a specific time point t_i, the basic formula (1) must be rewritten as follows,

$$A(p,i) = \frac{\pi_{p,i}\mu_{p,i}}{\sum_{j=1}^{n}\sum_{q\in V_j}\pi_{q,j}\mu_{q,j}},\ \forall p,i \text{ with } 1 \le i \le n, p \in V_i \tag{2}$$

where the total number of time points is n and V_j is the set of pages at time point t_j.

Finally to provide a temporal ranking model, the previously computed authority score is combined with a ranking function (e.g. Okapi BM25).

3 Strengths, Weaknesses, Extensions

In this section we discuss both the strong points and the weak aspects of the method presented in the paper. Besides noticing weaknesses, we have also included some extensions whose purpose is to improve or even to solve the issues associated with the respective weak point.

Accounting for freshness in the scoring function is an intuitively smart improvement for an information retrieval scoring model. Besides relevance and authority, freshness is an appropriate measure of the quality of a page. Pages being fresh tend to be welcome. However, traditional link analysis algorithms such as PageRank estimate page authority by simply accumulating contributions from in-links on a static web link structure, without considering whether pages are still fresh when web users search for them. Freshness of web links is also important to link-based ranking algorithms. The web is widely recognized as one of the networks in which the rich get richer as the networks grow, leading to power law effects. Old pages have more time to attract in-links, but may contain stale information. As an example consider the following situation: http://www.sigir2007.org/ has 902 in-links while http://www.sigir2010.org/ has only 208. Assuming the same contribution from each in-link, methods like PageRank would render a higher authority score on the earlier version of the SIGIR conference homepage. So, given this example, the advantages of the method presented in the paper are obvious.

However, the solution presented in the paper suffers from a number of shortcomings, presented in the following paragraphs.

One of the first weaknesses of the method is related to the manner in which the page freshness is propagated. More specifically, the authors propose a so-called backwards propagation which means that whenever a specific page gains freshness score all the pages out-linking this page will also be rewarded with a freshness score increase. In our opinion this technique is counter-intuitive as it promotes the idea of gaining without deserving and as a supporting argument we give the following example. Consider a page which was created in the 90's and which, for a reason or another, out-links a significant number of pages (e.g.

news.yahoo.com) which are very frequently updated (consequently, with a high page freshness) . In this case, the page freshness of the old and non-maintained page will be high and, depending on the linear combination of the relevance and authority/freshness components of the final score, it could take over pages which should be ranked better. Bottom line, the idea is that propagating the page freshness backwards rises issues regarding the fairness of the final score.

Another weak aspect of the paper is the estimation of the time the user spends on a particular web-page. The solution proposed in the paper is to estimate this time with the sum of the freshness of the in-links to the page under analysis. Besides being a little bit counter-intuitive we also think that this estimation is a little bit oversimplified and of a poor quality. Of course, at a first glance, we will observe that the more fresh the in-links of a page are the more likely the user is to reach that page and, by analogy to the web-surfer model augmented with the freshness theory, this assumption might be correct. But, this doesn't necessarily mean that the user is going to spend time there. It is very likely that the user will reach the page but it will not even take a look at it or it might leave it in a blink of an eye. Of course, we can also accept the fact that the authors are just trying to come up with an estimator and not with an exact measure of the time spent by the user on a page but as we've already mentioned this estimator could suffer from a large bias. Related to this weakness, an interesting extension would be to rely on the log files of the browser in order to derive a better estimator of the time spent by the user on a page.

Related to this weak aspect, we could also notice another downside of the method. More specifically, the distribution of the weights used the in the calculation of the time spent by the user on a particular page:

$$\mu_{p,i} = \sum_{t_j \in T'_{t_i}} \omega^{'}(t_i, t_j) InF(p)_{t_j} \tag{3}$$

is proposed to be an uniform distribution, i.e. all the snapshots are regarded to be equally important and this might not always be the case. We consider that there are far better distributions to be used in this case. Intuitively speaking, we should give more importance to more recent snapshots than to the older ones. Probably, distributions functions similar to the kernel functions used to model the authority propagation between snapshots would be better.

Yet another weakness of the method presented in the paper is represented by the sensitivity of the estimation of the page freshness to the number of detected page activities. Page freshness is given by the formula:

$$\Delta PF(q)|_{t_{i-1}}^{t_i} = \lambda_{PF}\, \Delta PF_0(q)|_{t_{i-1}}^{t_i} + (1 - \lambda_{PF}) \sum_{l:q \to p} m^{'}_{qp}\, \Delta PF(p)|_{t_{i-1}}^{t_i} \tag{4}$$

As we can see from the above formula, the freshness of a particular page is highly dependent on the number of page activities done on that page. To be more specific, the above formula is, most likely, an estimation of the page freshness (at

least this is what is going to be in a real world system) and the inexactness of the formula comes from the fact that we will never going to be able to know exactly which activities the page has undergone. As a matter of fact, this problem should be discussed in conjunction with another delicate related problem, i.e. the crawling strategy to be employed. As mentioned previously, the more page activities we detect the better the estimator will be; consequently, the crawler should be tuned to focus on the set of pages whose activity rate is larger; otherwise, employing a standard, uniform crawler might lead to a biased estimator. The very same argumentation applies also to link activity. Such a smart crawler would be a nice extension of the method, in case it's not already implemented. Nevertheless the authors do not discuss this aspect at all.

Another downside of the solution proposed in the paper is the performance aspect. Assuming that the method works smoothly and that it brings all the benefits of taking into consideration the freshness, there is still a noticeable shortcoming: the performance. The amount of resources needed by this method is significantly higher than those required by a standard authority computation algorithms, such as PageRank. To mention a few: Storage consumption is significantly increased by the need of permanently maintaining a number of snapshots. Obviously, a high number of snapshots implies better results of the method but also a huge demand for resources. In other words, the storage requirements are in correlation with the maximum size of the time window $|T|$. CPU resources are also in high demand in the case of this method – first of all, because we run PageRank on a number of instances of the Web Graph and not only on one instance as in the case of the standard Page Rank and, secondly, detecting the page and link activities requires a huge number of expensive string operations which are not needed in the standard PageRank. The sophistication of the crawler implies larger development times which is, sometimes, another scarce resource. To sum up, we can say that even if the method brings in some benefits, it does that at a high price.

Another downside of the method consists in the fact that each modification of a particular page triggers also the modification of the links inside that page (even though the respective links have not been modified). Besides the fact that we consider this as being counter-intuitive, there is also a more precise problem. Namely, no matter what small modifications we make to a page all its children will gain some in-link freshness without any reason. In the same context of link freshness, we also identified another negative point of the paper which even though it might seem insignificant and pertaining to the low-level details of the method, might rise some problems. Namely, one of the activities that can be done on a link is to modify any of its components excepting the anchor (see table 1). Given the formula of the in-link freshness:

$$\Delta InF(p)|_{t_{i-1}}^{t_i} = \lambda_{InF}\, \Delta InF_0(p)|_{t_{i-1}}^{t_i} + (1 - \lambda_{InF}) \sum_{l:q \to p} m_{qp}\, \Delta InF(q)|_{t_{i-1}}^{t_i} \quad (5)$$

we can observe that it will be affected by modifications such as color, style, or link-text whose purpose is not to reconfirm the freshness of the link but, by

the contrary, to mark it as obsolete or broken. As a concrete example, we can mention for instance: the modification of the text of the link to broken link or the modification of the color so that the link is not so visible anymore. Obviously, the real purpose of these actions is to diminish the importance of the link. Nevertheless, they are taken for proofs of in-link freshness.

Summing up, we can say that the paper presents a very useful method for improving authority scores (by considering the freshness) but the method suffers from a number of deficiencies which does not allow it to be regarded as a strong solution but more as just another algorithm in the large picture of similar methods.

References

[1] "Freshness Matters: In Flowers, Food, and Web Authority", Na Dai, Brian D. Davison, SIGIR 2010