Capturing Page Freshness for Web Search

Na Dai and Brian D. Davison
Department of Computer Science & Engineering
Lehigh University
Bethlehem, PA 18015 USA
{nad207,davison}@cse.lehigh.edu

ABSTRACT

Freshness has been increasingly realized by commercial search engines as an important criteria for measuring the quality of search results. However, most information retrieval methods focus on the relevance of page content to given queries without considering the recency issue. In this work, we mine page freshness from web user maintenance activities and incorporate this feature into web search. We first quantify how fresh the web is over time from two distinct perspectives—the page itself and its in-linked pages—and then exploit a temporal correlation between two types of freshness measures to quantify the confidence of page freshness. Results demonstrate page freshness can be better quantified when combining with temporal freshness correlation. Experiments on a real-world archival web corpus show that incorporating the combined page freshness into the searching process can improve ranking performance significantly on both relevance and freshness.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Performance

Keywords: temporal correlation, web freshness, web search

1 Introduction

Web search engines exploit a variety of evidence in ranking web pages to satisfy users' information needs as expressed by the submitted queries. These information needs may contain distinct implicit demands, such as relevance and diversity. Recency is another such need, and so is utilized as an important criteria in the measurement of search quality. However, most information retrieval methods only match queries based on lexical similarity. Link-based ranking algorithms such as PageRank [1] typically favor old pages since the authority scores are estimated based on a static web structure and old pages have more time to attract in-links.

To overcome this problem, we quantify page freshness from web activities over time. We observe that pages and links may have diverse update activity distributions from inception to deletion time points. We infer that pages having similar activity distributions with their in-links suggest that such page activities have stronger influence on their parents' activities.

Motivated by the above analysis, in this work we incorporate a temporal freshness correlation (TFC) component in quantifying page freshness, and show that by using TFC, we can achieve a good estimate of how up-to-date the page tends to be, which is helpful to improve search quality in terms of both result freshness and relevance. We consider the effects of other aspects of freshness on retrieval quality elsewhere [4].

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

	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 anch.)	↑	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.)

2 Page Freshness Estimation

We start by quantifying web freshness over time. We assign every page two types of freshness: (1) *page freshness* (PF) inferred from the activities on the page itself; and (2) *in-link freshness* (InF) inferred from the activities of in-links. Table 1 lists the detailed web activities and their contributions[1] to page and in-link freshness. To simplify analysis, we break the time axis into discrete time points $(t_0, t_1, \ldots, t_i, \ldots)$ with a unit time interval $\Delta t = t_i - t_{i-1}$, where $i > 0$. It is reasonable to assume that any activities that occur in $[t_{i-1}, t_i]$ can be considered as occurring at t_i, especially when Δt is small. We assume that the influence of activity decays exponentially over time. Therefore, we estimate PF and InF at t_i by aggregating the web activities with such a decay, written as:

$$PF_{t_i}(p) = \sum_{t_j=1}^{t_i} e^{(i-j)\alpha\Delta t} \sum_{k \in PA} w_k C_{t_j,k}(p)$$

$$InF_{t_i}(p) = \sum_{t_j=1}^{t_i} e^{(i-j)\alpha\Delta t} \sum_{l:q\rightarrow p} \sum_{k \in LA} w'_k C'_{t_j,k}(l)$$

where w_k and w'_k are contributions associated with each type of page and link activities, and $C_{t_j,k}(p)$ is the number of the k^{th} type of page activity on page p at t_j, and $C'_{t_j,k}(l)$ is the number of the k^{th} type of page activity on link l at t_j, and PA and LA are the page and link activity sets. In this way, we estimate web page freshness at multiple predefined time points from web activities.

[1]The sensitivity of activity weights with respect to freshness estimation is omitted due to space limitation.

We next quantify the temporal freshness correlation between pages and their in-links. We exploit the method by Chien and Immorlica [3], in which the authors measure query semantic similarity by using temporal correlation. Given a page p, its page and in-link freshness are denoted as $(PF_{t_c}(p), PF_{t_{c+1}}(p), \ldots, PF_{t_r}(p))$ and $(InF_{t_c}(p), InF_{t_{c+1}}(p), \ldots, InF_{t_r}(p))$ covering p's life span. The temporal freshness correlation (TFC) between page p and its in-links is given by:

$$TFC(p) = \frac{1}{n}\sum_{t=t_c}^{t_r}\left(\frac{PF_t(p) - \overline{PF(p)}}{\sigma_{PF}(p)}\right)\left(\frac{InF_t(p) - \overline{InF(p)}}{\sigma_{InF}(p)}\right)$$

where $\sigma_{PF}(p)$ and $\sigma_{InF}(p)$ are the standard deviations of $PF(p)$ and $InF(p)$, respectively.

Once we calculate the temporal freshness correlation for every page ($t_r - t_c \geq 2\Delta t$), we next combine it with page freshness score by ranks. Given a time point of interest t_i, the combined page freshness rank of document d is written as:

$$Rank_{combined}(d) = (1-\beta)Rank_{PF_{t_i}}(d) + \beta Rank_{TFC}(d)$$

where $\beta = \frac{a-1}{n-1+a-1}$, and n is the total number of time points, and a is the number of time points on which p exists. As a increases, $TFC(d)$ is more stable, and therefore we emphasize its contribution in the combined page freshness estimation.

3 Experimental Results and Discussion

Our goal is to improve web search quality on both relevance and freshness. To test the effect of combined page freshness on web search, we use an archival corpus of the `.ie` domain provided by the Internet Archive [5], covering from Jan. 2000 to Dec. 2007, and extract page and link activities. To minimize the influence of transient pages, we remove pages with fewer than 5 archival snapshots. The remaining sub-collection (with 3.8M unique URLs and 908M temporal links) is used for ranking evaluation.

We choose April 2007 as our time point of interest. 90 queries are selected from popular queries in Google Trends[2] for evaluation. For each query, we have an average of 84.6 URLs labeled by at least one worker of Amazon Mechanical Turk[3]. Editors give judgments on each document with respect to a given query for both relevance and freshness. Relevance is judged from "highly relevant" (4) to "not related" (0). Freshness is judged from "very fresh" (4) to "very stale" (0). The document with an average score above 2.5 is marked as relevant/fresh.

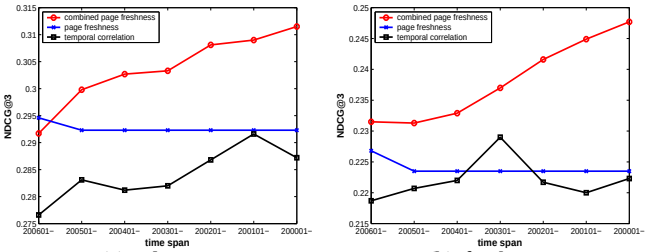
To evaluate the effectiveness of the combined page freshness, we compare with PageRank, running on a single web snapshot of April 2007. The global ranking lists generated by the combined page freshness and PageRank scores are linearly combined with Okapi BM2500 [6] (baseline) by ranks individually. The parameters are the same as Cai et al. [2]. Precision@k and NDCG@k are used as metrics for ranking evaluation on both relevance and freshness. All methods are compared based on their best rank combination of query-specific scores and global scores on metric Precision@10 of relevance. The decay parameter α is set to 1 in this work.

Table 2 lists the ranking performance comparison varying the time span involved in the combined page freshness computation. For relevance, except for NDCG@3, the correlation between ranking performance and the time span is not consistent. Unlike relevance, freshness performance consistently improves with the increase of time span used in the combined page freshness computation. This suggests temporal freshness correlation calculated from long-term web freshness measures can benefit more on accurate page freshness estimation. Figure 1 shows the performance on NDCG@3 with the variance of the time span for both relevance and freshness. We observe that (1) the ranking performance of page freshness first decreases, and then keeps nearly constant with the increase of time span, indicating the page activities within the past 1-2 years influence page freshness estimation the most; (2) the ranking performance of temporal freshness correlation shows unstable trends with variance of time span; and (3) the combined page freshness shows promising performance, and demonstrates its superiority over either page freshness or TFC.

[2]http://www.google.com/trends
[3]http://www.mturk.com

Relevance				
Method	P@10	NDCG@3	NDCG@5	NDCG@10
Okapi BM2500	0.4695	0.2478	0.2740	0.3344
PageRank	0.4894	0.2589	0.2840	0.3457
200601-200704	**0.5021**†	0.2917††	0.3152††	**0.3675**††
200401-200704	0.4893	0.3027††	0.3201††	0.3657††
200201-200704	0.5002†	0.3081††	0.3157††	0.3642††
200001-200704	0.4986†	**0.3115**††	**0.3211**††	0.3647††
Freshness				
Method	P@10	NDCG@3	NDCG@5	NDCG@10
Okapi BM2500	0.3138	0.2137	0.2379	0.2805
PageRank	0.3325	0.1946	0.2345	0.2838
200601-200704	0.3288†	0.2315††	0.2490†	0.2979†
200401-200704	0.3342†	0.2329††	0.2552††	0.2988†
200201-200704	0.3361†	0.2416††	0.2565††	0.3027††
200001-200704	**0.3374**†	**0.2477**††	**0.2617**††	**0.3028**††

Table 2: Ranking performance comparison. A † means the performance improvement is statistically significant (p-value<0.1) over Okapi BM2500. Performance improvement with p-value<0.05 is marked as ††.



(a) relevance (b) freshness

Figure 1: Ranking performance on metric NDCG@3 while varying the time span involved in page freshness calculation.

Acknowledgments

This work was supported in part by a grant from the National Science Foundation under award IIS-0803605 and an equipment grant from Sun Microsystems. We also thank Anlei Dong for helpful comments on the ranking evaluation criteria issue.

4 References

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proc. of 7th Int'l World Wide Web Conf.*, pages 107–117, Apr. 1998.

[2] D. Cai, X. He, J. Wen and W. Ma. Block-level link analysis. In *Proc. 27th Annual Int'l ACM SIGIR Conf.*, pages 440–447, Jul, 2004.

[3] S. Chien and N. Immorlica. Semantic similarity between search engine queries using temporal correlation. In *Proc. 14th Int'l World Wide Web Conf.*, pages 2–11, 2005.

[4] N. Dai and B. D. Davison. Freshness Matters: In Flowers, Food, and Web Authority. In *Proc. of 33rd Annual Int'l ACM SIGIR Conf.*, Jul, 2010.

[5] The Internet Archive, 2010. http://www.archive.org/.

[6] S. E. Robertson. Overview of the OKAPI projects. *Journal of Documentation*, 53:3–7, 1997.