Freshness Matters: In Flowers, Food, and Web Authority

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

ABSTRACT

The collective contributions of billions of users across the globe each day result in an ever-changing web. In verticals like news and real-time search, recency is an obvious significant factor for ranking. However, traditional link-based web ranking algorithms typically run on a single web snapshot without concern for user activities associated with the dynamics of web pages and links. Therefore, a stale page popular many years ago may still achieve a high authority score due to its accumulated in-links. To remedy this situation, we propose a temporal web link-based ranking scheme, which incorporates features from historical author activities. We quantify web page freshness over time from page and in-link activity, and design a web surfer model that incorporates web freshness, based on a temporal web graph composed of multiple web snapshots at different time points. It includes authority propagation among snapshots, enabling link structures at distinct time points to influence each other when estimating web page authority. Experiments on a real-world archival web corpus show our approach improves upon PageRank in both relevance and freshness of the search results.

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

General Terms: Algorithms, Performance

Keywords: Web search engine, temporal link analysis, web freshness, PageRank

1 Introduction

In a corpus of documents as large as the Web, relevance alone is often insufficient to produce good rankings. Thus it is necessary to differentiate pages, and for more than a decade search engines have considered page authority in addition to relevance with respect to queries (and other factors) in web search. Much previous work [10, 23, 25] has been studied to estimate page authority based on different assumptions and successfully generalized onto multiple tasks [5, 9, 32]. However, most of these studies accumulated the authority contributions only based on the evidence of links between pages, without considering the temporal aspects concealed in pages and their connections.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

Freshness is important to the quality of much in our daily lives, such as flowers and food. The same is also true for web page authority estimation. Pages being fresh tend to be welcome. However, traditional link analysis algorithms such as PageRank [10] 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 [12]. Old pages have more time to attract in-links, but may contain stale information. For example, as of this writing, `http://www.sigir2007.org/` has 902 in-links [33] while `http://www.sigir2010.org/` only has 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.

In addition, some web spam research [31, 14] has recognized that local link structures with sudden changes might indicate link spam. A single web snapshot is unable to detect such changes and further smooth or neutralize the undesirable influence automatically.

Motivated by these two points, in this work we propose an probabilistic algorithm to estimate web page authority by considering two temporal aspects. First, to avoid old pages from dominating the authority scores, we keep track of web freshness over time from two perspectives: (1) how fresh the page content is, referred to as *page freshness*; and (2) how much other pages care about the target page, referred as *in-link freshness*. To achieve this, we mine web authors' maintenance activities on page content, such as the creation and removal of out-links. Each activity is associated with the time at which it occurs. We build up temporal profiles for both pages and links. A random walk model is exploited to estimate the two predefined freshness measures. By modeling web freshness from these two perspectives, we can bias the authority distribution to fresh pages, and so neutralize the unfair preference toward old pages by traditional link analysis ranking algorithms.

Second, we use multiple web snapshots at distinct time points, instead of a single snapshot. To make the link graph more stable, we connect multiple web snapshots by propagating authority flows among them, and so smooth the impact of sudden changes to particular snapshots on web page authority estimation. We exploit several proximity-based density kernel functions to model such propagation. Combining web freshness measures, we utilize a semi-Markov process to model a web surfer's behavior in selecting and browsing web pages.

We show the superiority of our proposed approach by conducting experiments to evaluate the ranking performance of several representative temporal web link-based ranking algorithms on a real-

world archival web data set. Experimental results demonstrate that our method outperforms the original time-agnostic PageRank by 8% on both relevance and freshness of top 10 search results.

The contributions of this work are as follows.

- Quantify web freshness from authors' maintenance activities on web content over time, from the perspectives of page freshness and in-link freshness;
- Incorporate web freshness into authority propagation to favor fresh pages;
- Explore a series of proximity-based density kernel functions to model authority propagation among web snapshots;
- Conduct experiments on a real-world archival web data set and show the superiority of our approach on ranking performance in terms of both relevance and freshness.

The remainder of this paper is organized as follows. We review prior work in Section 2; introduce how we quantify web freshness and how we incorporate it into a web surfer model to estimate time-dependent web page authorities in Section 3; present how we set up experiments in Section 4; and show the evaluation results of our proposed ranking algorithm in Section 5. We conclude and consider future work in Section 6.

2 Related Work

Link analysis which incorporates temporal aspects into search has been studied in [35, 4, 13, 8, 6, 7, 34, 24, 1]. Yu et al.'s work in [35] was among the earliest, which incorporated paper age into quantifying paper authority to improve academic search. In addition to utilizing paper citations, the authors modified PageRank by weighting each citation according to the citation date. However, their work only associated one type of activity, i.e., link (citation) creation, into link analysis in the scenario of academic search. Amitay et al. [4] attached a timestamp to each link, approximating the age of the page's content. However, they only gave bonus to the links from fresh pages, rather than combining the freshness of the page itself when estimate web page authority. Berberich et al.'s work [8] focused on temporal aspects of both web pages and links in web search via the web dynamics from page and link creation, modification and deletion. They assumed users are equally interested in recency of information, in addition to the quality. Their work specifically emphasized the freshness and activity of pages and links. However, activities occurring at different time points are not distinguished as long as they were all in the period of users' temporal interests, which could span wide ranges.

Our work differs from prior work in two ways. First, we model the web freshness from two different perspectives by building temporal link profiles and temporal page profiles from multiple types of activities over time. Second, the influence of activities on web freshness decays over time. We include some of these methods (e.g., [35, 8]) for comparison in Section 5.

Another direction in link analysis which incorporates temporal factors is to directly utilize or mine trends from multiple snapshots of the archival web [6, 7, 34, 24]. Berberich et al. [6] analyzed the potential of page authority by fitting an exponential model of page authority. The success with which web pages attract in-links from others in a given period becomes an indicator of the page authority in the future. However, one problem is how to normalize page authority at different time points, such that they are comparable. To solve it, Berberich et al. [7] normalized PageRank scores by dividing them by the minimum authority score in the same web snapshot, so that the minimum normalized PageRank score of the page in any snapshot equals 1.

Yang et al. [34] proposed a new framework which utilizes a kinetic model to explain the evolution of page authority over time from a physical point of view. The page authorities are viewed as objects subject to both "driving force" and "resistance", and so the page authority at any time point can be a combination of the current authority score resulting from "driving force" and the decayed historical authority score from "resistance". Empirical experiments demonstrated that authority estimation can benefit from increasing use of archival web content. However, their method did not consider the accumulation of incomparable authority scores caused by the inconsistent numbers of the pages in distinct snapshots. Other than web search, the idea of propagation of authority flows among different snapshots has been found in some other domains, such as social network analysis. Li and Tang [24] modeled the decayed effects of old publications in expertise search by allowing authority exchange only between successive snapshots of the time-varying social networks.

Our work differs from these in two ways. First, in our method each page in any snapshot is directly influenced by the same page in all the snapshots in a one-step transition decayed by time difference. This process captures a comprehensive interaction between pages at different time points naturally. Second, we propose and evaluate a series of proximity-based kernel functions to control the authority propagation among multiple snapshots. Again, we compare to some of these approaches (e.g., [6, 34, 24]) in Section 5.

Many link analysis methods compute page authority by a stochastic process via the link structure of the web. However, Liu et al. [25] utilized users' browsing behaviors to calculate page authority from a continuous-time Markov process which combines both how likely a web surfer reaches a page and how long the web surfer stays on a page. Their follow-up work [17] generalizes the page importance framework to be a semi-Markov process in which how long a web surfer stays on a page can partially depend on where the surfer comes from in one step transition. Since our work models web freshness from both how fresh a page is and how well other pages care about a particular page over time, we incorporate these two aspects into a semi-Markov process, which can model a temporal web surfer behavior in a natural and adaptive way. Other related work includes exploring temporal aspects of web behaviors [2] and utilizing the evolution patterns of pages and links to benefit web-based applications [16].

3 Methodology

In this section, we propose a temporal ranking model (denoted as T-Fresh) to incorporate web freshness and link structures at different time points into web page authority estimation. The main idea of T-Fresh is to utilize authors' maintenance activities on web pages and links to estimate web freshness at different time points, and then incorporate them into time-dependent page authority estimation by following proximity-based authority propagation rules on the time axis. T-Fresh outputs an authority score for each page at every predefined time point. The authority is estimated in an approximated way, partly depending on the link structure and web freshness of nearby snapshots, with the ones at farther time points having smaller influence.

3.1 Representing Web Freshness Over Time

As introduced in Section 1, web freshness reflects how fresh a web page is at a given time point t_i by in-link freshness (InF) and page freshness (PF). The reasons we separate these two web freshness measures are: (1) InF and PF depict web freshness from the perspectives of information recommenders and information providers respectively; and (2) it prevents one type of web freshness from

	Link activity	Infl. on p's InF	Gain of p's InF
1	creation of link $l: q \rightarrow p$	$\uparrow\uparrow\uparrow$	3
2	update on link $l: q \rightarrow p$ (changed anchor)	$\uparrow\uparrow$	2
3	update on link $l: q \rightarrow p$ (unchanged anchor)	$\uparrow$	1.5
4	removal of link $l: q \rightarrow p$	$\downarrow\downarrow$	-0.5
	Page activity	**Infl. on q's PF**	**Gain of q's PF**
1	creation of page q	$\uparrow\uparrow\uparrow$	3
2	update on page q	$\uparrow$	1.5
3	removal of page q	$\downarrow\downarrow$	-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. $\uparrow$: positive influence on web freshness. $\downarrow$: negative influence on web freshness. The number of $\uparrow$ or $\downarrow$ indicates the magnitude.)

dominating a single freshness score. Given a web page p, we assume that each update on p's parent page q is a direct validation of the link from q to p, and so an update on q implies that q pays attention to all of its out-linked pages, including p. Hence, we use InF to represent the attention from p's in-link pages, which is computed from the accumulation of activities on all the p's parent pages up to t_i. Unlike InF, PF represents how fresh p is up to t_i based on the activities on page p itself. For every page p at t_i, it associates with InF and PF, denoted as $InF(p)_{t_i}$ and $PF(p)_{t_i}$.

3.1.1 Building Temporal Page and Link Profiles

In order to compute InF and PF, we generate temporal page profiles (TPP) and temporal link profiles (TLP) in a manner inspired by Amitay et al. [4]. TPP and TLP record the web authors' activities on the pages and links over time. Given a page p, each item on its TPP records evidence of some type of activity on p at a specific time point. It is written as a 3-tuple <page ID, activity type, timestamp>, where activity type∈{creation, update, removal}. Given a link l with its associated anchortext, TLP records the evidence of some type of activity on l at a specific time point. Each item on TLP can similarly be represented as the 3-tuple <link ID, activity type, timestamp>, where activity type∈{creation, update with unchanged anchor, update with changed anchor, removal}. In this way, each link and page is associated with a series of timestamped activities. Table 1 summarizes the influence of these activities on web freshness.

3.1.2 Quantifying Web Freshness

Based on TPP and TLP, we next quantify web freshness, i.e., InF and PF. In order to simplify analysis, we separate the continuous time axis into discrete time points, e.g., $(t_0, t_1, \ldots, t_n, \ldots)$, with a unit time interval Δt between successive time points, i.e., $\Delta t = t_i - t_{i-1}$. Web freshness at any time point t_i is dependent on (1) the web freshness at t_{i-1}, and (2) the activities on TPP and TLP, which occur between t_{i-1} and t_i. When Δt is small enough, it is reasonable to assume that any activities in $[t_{i-1}, t_i]$ occur at t_i. In this way, we map all the web activities onto discrete time points. For web freshness at t_{i-1}, we assume it decays exponentially over time. Thus, $InF(p)_{t_i}$ and $PF(p)_{t_i}$ can be given by:

$$InF(p)_{t_i} = \beta_1 e^{-\beta_2 \Delta t} InF(p)_{t_{i-1}} + \Delta InF(p)|_{t_{i-1}}^{t_i} \quad (1)$$

$$PF(p)_{t_i} = \beta_3 e^{-\beta_4 \Delta t} PF(p)_{t_{i-1}} + \Delta PF(p)|_{t_{i-1}}^{t_i} \quad (2)$$

where $\Delta PF(p)|_{t_{i-1}}^{t_i}$ and $\Delta InF(p)|_{t_{i-1}}^{t_i}$ are the incremental freshness scores from the activities in $[t_{i-1}, t_i]$, and $\beta_1 e^{-\beta_2 \Delta t}$ is a coefficient that controls the decay of historical web freshness.

In the next step, we compute the incremental in-link freshness $\Delta InF(p)|_{t_{i-1}}^{t_i}$ for the given page p. Since in-link freshness depends on the activities on TLP, we compute $\Delta InF(p)|_{t_{i-1}}^{t_i}$ by accumulating all the activities on p's in-links in $[t_{i-1}, t_i]$. Let $C_j(l)$ be the number of the j^{th} type of link activity on link l in $[t_{i-1}, t_i]$. Let w_j be the unit contribution of the j^{th} type of link activity. The incremental in-link freshness is written as:

$$\Delta InF_0(p)|_{t_{i-1}}^{t_i} = \sum_{l:q\rightarrow p} \sum_{j \in LA} w_j C_j(l) \quad (3)$$

where LA is the set of link activity types. However, it is not enough to propagate such influence in one step; we additionally propagate in-link activities in an iterative way, leading to smoother in-link freshness scores. Let $\Delta InF_0(p)|_{t_{i-1}}^{t_i}$ in Equation 3 be an initial score. For each iteration, every page receives in-link freshness scores from its parent pages, and also holds its initial score. The process converges and produces a stable score for every page determined by both its parents' scores and its own in-link activities. Thus, the incremental in-link freshness is given by:

$$\begin{aligned} \Delta InF(p)|_{t_{i-1}}^{t_i} &= \lambda_{InF} \Delta InF_0(p)|_{t_{i-1}}^{t_i} \\ &\quad + (1-\lambda_{InF}) \sum_{l:q\rightarrow p} m_{qp} \Delta InF(q)|_{t_{i-1}}^{t_i} \end{aligned} \quad (4)$$

where m_{qp} is the weight on the link from q to p. Equation 4 is actually the personalized PageRank (PPR) [19]. We use one-step transition probability from q to p based on link structure to represent m_{qp}, where $\sum m_{q*} = 1$ if q has at least one out-link.

We next compute the incremental page freshness $\Delta PF(p)|_{t_{i-1}}^{t_i}$. Similar to $\Delta InF(p)|_{t_{i-1}}^{t_i}$, we argue that how fresh one page is depends on both the page itself and its out-linked pages, since the out-linked pages are extensions of the current page. We thus propagate page freshness backward through links in an iterative way. For each iteration, every page receives page freshness scores from its out-linked pages, and also holds its initial score. This process converges finally and generates a stable page freshness score on every page. Let $C'_j(p)$ be the number of the j^{th} type of page activity on p in time period $[t_{i-1}, t_i]$. Let w'_j be the unit contribution of the j^{th} type of page activity. The initial incremental page freshness score $PF_0(p)|_{t_{i-1}}^{t_i}$ is defined as:

$$\Delta PF_0(p)|_{t_{i-1}}^{t_i} = \sum_{j \in PA} w'_j C'_j(p) \quad (5)$$

where PA is the set of page activity types. The incremental page freshness is given by:

$$\begin{aligned} \Delta PF(q)|_{t_{i-1}}^{t_i} &= \lambda_{PF} \Delta PF_0(q)|_{t_{i-1}}^{t_i} \\ &\quad + (1-\lambda_{PF}) \sum_{l:q\rightarrow p} m'_{qp} \Delta PF(p)|_{t_{i-1}}^{t_i} \end{aligned} \quad (6)$$

where m'_{qp} is the weight on the link from q to p. We use the inverted one-step transition probability to represent m'_{qp}, where $\sum m'_{*p} = 1$ if page p has at least one in-link. Once achieving $\Delta InF(p)|_{t_{i-1}}^{t_i}$ and $\Delta PF(p)|_{t_{i-1}}^{t_i}$, we compute $InF(p)_{t_i}$ and $PF(p)_{t_i}$ by Equation 1 and 2.

3.2 Temporal Ranking Model

Now that we quantify web freshness at distinct time points, the next problem comes to how to utilize web freshness to control authority propagation in an archival link graph, so that we achieve a time-dependent authority score for every page.

We start by describing a "temporal random surfer model" which motivates our method T-Fresh. The "temporal random surfer

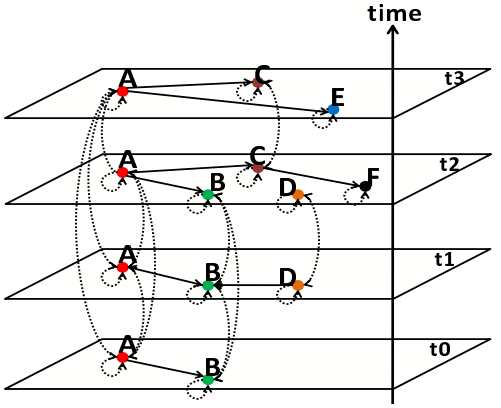


Figure 1: The process of T-Fresh. Each node represents one web page.

model" is similar to the "random surfer model", which explains PageRank [10]. However, our surfer model differs from the traditional model in two aspects. First, the web surfer has specific temporal intent, which controls her choice of target snapshot. Note that the surfer's temporal intent varies with her current snapshot. Second, the web surfer prefers fresh web resources. Figure 1 depicts one simple example of how the surfer behaves on an archival web of four snapshots.

Consider a web surfer wandering on an archival web corpus, which includes multiple web snapshots collected at different time points (t_0, t_1, ..., t_n). For every move, the surfer takes the following steps. First, she can choose either to follow one of the out-linked pages or to randomly jump to any page at the same time point. However, unlike PageRank in which a web surfer has equal probabilities to follow out-going links, the preference of our surfer choosing out-going links is a function of the page freshness of out-linked pages. Consider the example in Figure 1. Suppose the surfer is currently on page A at t_2. She follows the link to B at t_2 (solid link) with probability $(1-d)F_{t_2}(B,A)$, where $F_{t_2}(B,A)$ is a function which depends on the page freshness of all A's out-linked pages at t_2 and $\sum_{P:A\to P} F_{t_2}(P,A)=1$. The probability that the surfer randomly jumps to any page within snapshot t_2, such as B, is d/N_{t_2}, where N_{t_2} is the total number of pages at t_2.

After the surfer reaches the page chosen in the first step, she next selects the specific snapshot of the page to which to jump based on her temporal intent, which correlates to the time difference between the current snapshot and the target snapshot. This process propagates authority among snapshots and uses the link structure at one time point to influence the authority computation at other time points. The propagation decays with time difference between snapshots. For the example in Figure 1 (dashed bi-directed links), suppose the surfer reaches B at t_2 after the first step, she can jump to B at any time point as long as it exists, i.e., t_2, t_1, and t_0. Specifically, the probability that she jumps to B at t_1 is written as $P_{t_1|t_2}(B)$, which depends on the time difference between t_1 and t_2.

Once the surfer reaches the page at the chosen time point, e.g., page B at t_1, she browses it with the mean stay time $\mu_{t_1}(B)$, which incorporates B's in-link freshness at t_1 before the next move.

In this way, the surfer's behavior on the archival web can be separated as (1) moving from one page to another; and (2) staying on a page and browsing it. It leads to the semi-Markov process [30] for page authority estimation.

Definition 1. A semi-Markov process is defined as a process that can be in any one of N states 1, 2, ..., N, and each time it enters a state i it remains there for a random amount of time having mean μ_i, and then makes a transition into state j with probability P_{ij}.

Suppose the time that the process spends on each state is 1; then the semi-Markov process leads to a Markov chain. Assuming all states in such a Markov chain communicate with each other, the process can generate a stationary probability π_i for any state i. The long-run proportion of time that the original semi-Markov process is in state i is given by:

$$A(i)=\frac{\pi_i\mu_i}{\sum_{j=1}^{N}\pi_j\mu_j}, i=1,2,\ldots,N \tag{7}$$

This solution divides the time-dependent page authority estimation into (1) computing the stationary probability that a surfer reaches every page in the archival corpus; and (2) computing the mean time of a surfer staying on every page.

3.2.1 Estimating Stationary Probability

We now introduce the computation of probability π_{p,t_i} that a web surfer enters a page p at snapshot t_i. In the first step of each move, the surfer reaches page p at any time point t_j by: (1) following p's in-link at t_j to reach p; (2) jumping from any page at t_j to p at t_j.

$$P_{t_j}(Follow|q)=(1-d), \quad P_{t_j}(p|q,Follow)=F_{t_j}(p,q) \tag{8}$$

$$P_{t_j}(Jump|q)=d, \qquad P_{t_j}(p|q,Jump)=1/N_{t_j} \tag{9}$$

where d is 0.15 by default. $F_{t_j}(p,q)$ is the web surfer's preference for following out-linked pages. Intuitively, fresh web resources are likely to attract surfer's attention. We define $F_{t_j}(p,q)$ as:

$$F_{t_j}(p,q)=\frac{PF_{t_j}(p)}{\sum_{p':q\to p'|_{t_j}}PF_{t_j}(p')} \tag{10}$$

In the second step of each move, the surfer reaches page p at t_i from page p at t_j is given by:

$$P_{t_i|t_j}(p)=\frac{w(t_i,t_j)}{\sum_{q\in V_i,q\in V_j}w(t_i,t_j)} \tag{11}$$

where V_i and V_j are the sets of pages at time point t_i and t_j respectively, and $w(t_i,t_j)$ is the weight that represents the influence between the snapshots at t_i and t_j. Motivated by previous work [15, 22, 26, 28] which used proximity-based methods, we consider six kernel functions to model the authority propagation between snapshots: gaussian kernel (equation 12), triangle kernel (equation 13), cosine kernel (equation 14), circle kernel (equation 15), passage kernel (equation 16) and PageRank kernel (equation 17). We formally define them as follows.

$$w_1(t_i,t_j)=\exp\left[-\frac{(t_i-t_j)^2}{2|T|^2}\right] \tag{12}$$

$$w_2(t_i,t_j)=1-\frac{|t_i-t_j|}{|T|} \tag{13}$$

$$w_3(t_i,t_j)=\frac{1}{2}\left[1+\cos\left(\frac{|t_i-t_j|\pi}{|T|}\right)\right] \tag{14}$$

$$w_4(t_i,t_j)=\sqrt{1-\left(\frac{|t_i-t_j|}{|T|}\right)^2} \tag{15}$$

$$w_5(t_i,t_j)=1 \tag{16}$$

$$w_6(t_i, t_j) = \begin{cases} 0.85 & t_i = t_j \\ \frac{0.15}{|T-1|} & t_i \neq t_j \end{cases} \quad (17)$$

where $|T|$ is the window size of one step authority propagation between snapshots. Other than Equation 12, all kernels require $|t_i - t_j| < |T|$; that is, the one step authority propagation proceeds only within a window of a specified size. Larger $|T|$ results in more choices for the web surfer at each move between snapshots, while smaller $|T|$ leads to influence mainly from nearby time points. In this work we set $|T|$ to the total number of snapshots involved in authority propagation by default.

Combining the analysis above, the probability that a web surfer reaches page p at snapshot t_i can be written as:

$$\begin{aligned} \pi_{p,i} &= \sum_{t_j \in T_i} P_{t_i|t_j}(p) \sum_{q:q\to p|_{t_j}} P_{t_j}(Follow|q) P_{t_j}(p|q, Follow) \\ &\quad + \sum_{t_j \in T_i} P_{t_i|t_j}(p) \sum_{q|_{t_j}} P_{t_j}(Jump|q) P_{t_j}(p|q, Jump) \\ &= \sum_{t_j \in T_i} P_{t_i|t_j}(p) \\ &\quad \times \left[(1-d) \sum_{q:q\to p|_{t_j}} F'_{t_j}(p,q)\pi_{q,j} + d \sum_{q|_{t_j}} \frac{\pi_{q,j}}{N_{t_j}} \right] \end{aligned} \quad (18)$$

where T_i is the set of snapshots which can directly distribute authority to t_i within one step. Based on the surfer's behavior, this Markov process guarantees all states to communicate with each other, leading to a transition matrix that is irreducible and aperiodic [30]. As a result, it converges and generates a stationary probability on every page existing in any snapshot.

3.2.2 Estimating Staying Time

Pages with more in-link activity are likely to attract a surfer to spend time browsing it. We assume the web surfer prefers fresh web resources, and so the mean time ($\mu_{p,i}$) of the surfer staying on page p at t_i can be proportional to p's web freshness at t_i. As discussed in Section 3.2.1, the web surfer prefers pages with high page freshness when choosing among out-going links; we use in-link freshness to model the time of a surfer staying on a web page. In this way, pages with both high in-link freshness and page freshness are more likely to be given high authority scores. Specifically, we utilize a sliding window and compute p's weighted in-link freshness centroid within it as the estimation of $\mu_{p,i}$, which is formally given by

$$\mu_{p,i} = k \sum_{t_j \in T'_{t_i}} w'(t_i, t_j) InF(p)_{t_j} \quad (19)$$

where T'_{t_i} is the set of snapshots included in the sliding window centered on t_i, and $\sum_{t_j \in T'_{t_i}} w'(t_i, t_j) = 1$. In this work we evaluate one special case, in which $w'(t_i, t_j) = \frac{1}{|T'_{t_i}|}$ for any $t_j \in T'_{t_i}$. In this way, the authority score $A(i)$ in Equation 7 is determined by both $\pi_{p,i}$ in Equation 18 and $\mu_{p,i}$ in Equation 19.

4 Experimental Setup

4.1 Data set and Evaluation

Many standard data sets such as TREC [27] usually only contain one snapshot of a web corpus, and so are not suitable to show the effectiveness of ranking models utilizing temporal information. To evaluate our proposed method, we utilize a corpus of archival web pages in the `.ie` domain collected by Internet Archive [20] from January 2000 to December 2007. This corpus contains 158 million unique web pages, and approximately 12 billion temporal links. To avoid the influence of transient web pages, we extract one web graph for each month from the sub-collection of pages for which we have at least 5 crawled copies. These graphs comprise a collection of 3.8M unique pages and 435M temporal links in total.

Notation of T-Fresh variants: T-Fresh(kernel, window, snapshot)	
kernel	The kernel controlling authority propagation among different web snapshots, where $kernel \in \{1, 2, 3, 4, 5, 6\}$
window	The window size used in calculating average in-link freshness for estimating staying time, where $window \in N$
snapshot	The number of months spanned over the temporal graph where $1 \leq snapshot \leq 88$ (from Jan. 2000 to Apr. 2007)

Table 2: Notation of T-Fresh variants.

For ranking evaluation, we choose `April 2007` as our time period of interest. Ninety queries are selected from a set of sources, including those frequently used by previous researchers, and popular queries from Google Trends [18]. For each query, we have an average of 84.6 URLs judged by at least one worker of Amazon's Mechanical Turk [3]. When human editors judge each pair of <query,URL>, they are required to give a score based on (1) how relevant the page is to the query; and (2) how fresh the page would be as a result for the requested time period. The relevance score is selected from among *highly relevant, relevant, borderline, not relevant* and *not related*, which is translated to an integer gain from 4 to 0. A page with score higher than 2.5 is marked as relevant. Similar to the relevance judgement, the freshness score is selected from *very fresh, fresh, borderline, stale,* and *very stale*, which we translate into an integer scaled from 4 to 0. A page with a score higher than 2.5 is marked as fresh. All human editors were asked to give the confidence of their provided judgments, in the selection of high, medium and low. Judgements with low confidence are not included in ranking evaluation. A random sample with 76 <query, URL> pairs judged by 3 editors show that the average standard deviations of relevance and freshness judgements are 0.88 and 1.02 respectively.

Based on these judgements, we evaluate the ranking quality of our approach on both relevance and freshness over the Normalized Discounted Cumulative Gain (NDCG) [21] metric. It penalizes the highly relevant or fresh documents appearing at lower positions. Precision@k is also utilized to measure the ranking quality, which calculates the number of relevant or fresh documents within the top k results across all queries.

4.2 Compared Methods

To show the effectiveness of T-Fresh, we compare with PageRank [10] (the baseline) and several representative link-based ranking algorithms, which incorporate temporal information, including TimedPageRank [35], T-Rank [8], BuzzRank [6], TemporalRank [34], and T-Random [24]. All these algorithms combine with Okapi BM2500 [29] linearly by ranks, defined as:

$$(1-\gamma) rank_{authority}(p) + \gamma rank_{BM}(p)$$

The parameters used in Okapi BM2500 are the same as Cai et al. [11]. The variants of T-Fresh are summarized in Table 2.

4.3 Web Activity Detection

While accurate web maintenance activities might be recorded on Web servers' logs, we must infer such activities from the comparison between successive web snapshots in this work. Specifically, we assume that each page was created at the time at which it was first crawled, and each link was created when it was first found. Although some pages can automatically change a portion of its con-

tent in every crawl, we suppose one page has an update when its content has any difference from the previous version, or its meta-data can show the last-modified time is after the crawling time of the previous one. To identify the link update, we simply assume that once a page has an update, all its out-links are considered to be updated. We admit that perfect quantification of link update activity may depend on a variety of factors, including the distance to page blocks being changed, page editing concentration inferred from content maintenance patterns, and so on. We leave the sensitivity of web activity detection accuracy on ranking performance to future work. We also assume that a page disappears when its returned HTTP response code is 4xx or 5xx. While the gain associated with each type of link and page activity can influence the ranking performance, as a preliminary study, we define these gains in Table 1, and again leave the sensitivity of ranking performance with respect to gains on web activity to future work.

5 Experimental Results

In this section, we report the results of our ranking evaluation and compare T-Fresh to representative link-based algorithms. Results demonstrate that by incorporating web freshness and propagating authority among different web snapshots, we can achieve more relevant and fresh search results.

5.1 Correlation of InF and PF

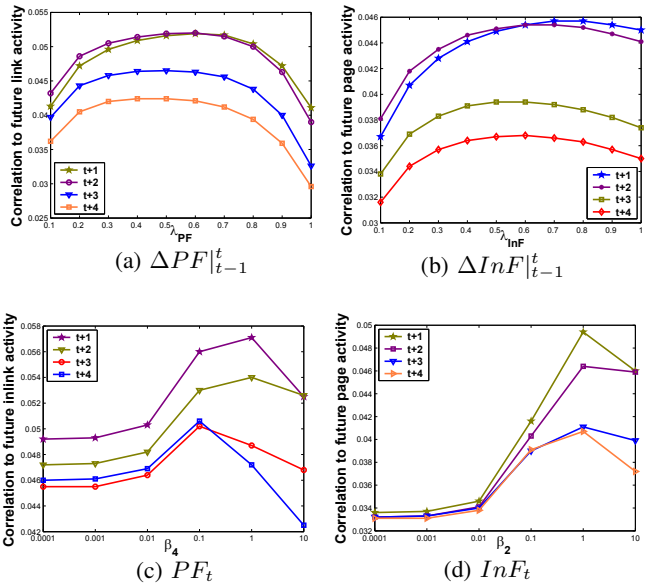
As introduced in Section 3.1.2, each page in the temporal graph is associated with InF and PF. A reasonable criteria for the good estimation of InF and PF would be their potential capability of predicting future web activities even though the correlation between them would be rather small. To better consider this idea, we compute the average correlation between web freshness scores at t and web activities at future time points, i.e., $t+1$, $t+2$, etc., given by Equation 3 and 5.

From Figure 2(a), $\Delta PF|_{t-1}^{t}$ and future in-link activity show positive correlation, with the strength inversely proportional to the time difference between the incremental page freshness and future in-link activities. In most cases, the correlation is maximized when λ_{PF} and λ_{InF} are 0.6. It indicates pages can achieve freshness scores from both activities on themselves and their neighbor pages via propagation. The correlations between $\Delta InF|_{t-1}^{t}$ and future page activities show similar trends (Figure 2(b)). One may notice that the average correlation between $\Delta PF|_{t-1}^{t}$ and in-link activities at $t+1$ is 0.0519, which is higher than that between $\Delta InF|_{t-1}^{t}$ and page activities at $t+1$ over 13.5%. One interpretation is that a page with very fresh content tends to attract new in-links or existing in-links to validate in next time periods. From Figure 2(c) and (d), the cumulative web freshness scores can show stronger correlation to future web activities, varying with the decay parameter β_2 and β_4 given $\beta_1 = \beta_3 = 1$ constantly. For both PF_t and InF_t, the correlations achieve the highest when β_2 and β_4 are 1 in most cases. To meet our criteria of good estimation about web freshness, we set $\lambda_{PF} = \lambda_{InF} = 0.6$ and $\beta_2 = \beta_4 = 1$ in the following ranking evaluation.

5.2 Ranking Performance

Figure 3 demonstrates the ranking performance in terms of relevance and freshness on metric P@10 over all the compared algorithms, under the variance of combination parameter γ from 0.8 to 1. The variant of T-Fresh we choose for comparison is T-Fresh(1,1,30). For relevance evaluation, PageRank achieves its highest P@10 at 0.4894 when γ is 0.97. T-Fresh performs the best among all the algorithms, achieving its highest P@10 at 0.5051 when γ is 0.91, exceeding PageRank by 3.2%. The TimedPageRank places second on metric P@10, which reaches 0.5031 when γ is 0.92. Under the combination parameter achieving the best P@10 for every method, we compare the ranking performance over all metrics in Table 3. T-Fresh performs the best among all the algorithms over all the metrics. Specifically, it outperforms PageRank over 24.7%, 17.8% and 7.8%, in terms of NDCG@3, NDCG@5 and NDCG@10. Single-tailed student t-tests at a confidence level of 95% demonstrate the improvements are statistically significant over PageRank on NDCG@3, NDCG@5 and NDCG@10, with p-values 0.0001, 0.0001, 0.0016 respectively.

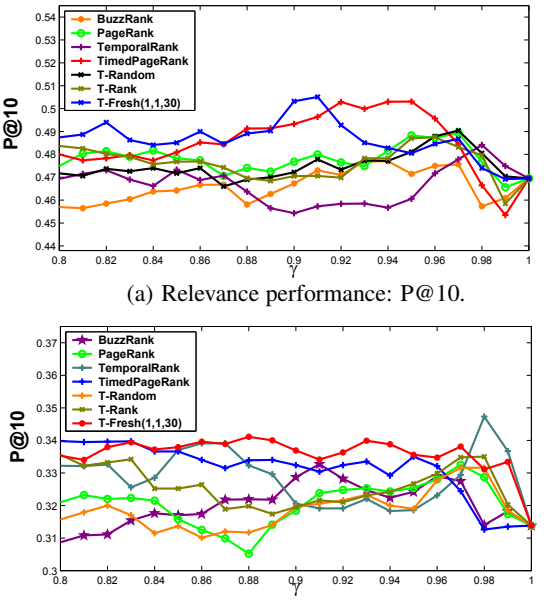
For freshness evaluation, Figure 3(b) shows ranking performance on metric P@10, varying with combination parameter γ. T-Fresh demonstrates a stable trend for P@10, which exceeds PageRank on all the experimental points. Unlike relevance evaluation in which improvements of other temporal link-based algorithms are



(a) $\Delta PF|_{t-1}^{t}$ (b) $\Delta InF|_{t-1}^{t}$

(c) PF_t (d) InF_t

Figure 2: Correlation between web freshness and future web activities.



(a) Relevance performance: P@10.

(b) Freshness performance: P@10.

Figure 3: Sensitivity of P@10 with respect to combination parameter γ.

Relevance				
Method	P@10	NDCG@3	NDCG@5	NDCG@10
BM25	0.4695	0.2478	0.2740	0.3344
PageRank	0.4894	0.2589	0.2840	0.3457
BuzzRank	0.4770	0.2770	0.2980	0.3460
TemporalRank	0.4841	0.2706	0.2875	0.3524
TimedPageRank	0.5031	0.2830	0.3063	0.3587
T-Random	0.4904	0.2690	0.2877	0.3495
T-rank	0.4875	0.2669	0.2870	0.3496
T-Fresh(1,1,30)	**0.5051**	**0.3229**	**0.3347**	**0.3729**
Freshness				
Method	P@10	NDCG@3	NDCG@5	NDCG@10
BM25	0.3138	0.2137	0.2379	0.2805
PageRank	0.3325	0.1946	0.2345	0.2838
BuzzRank	0.3327	0.2043	0.2234	0.2797
TemporalRank	**0.3473**	0.2312	0.2510	0.2992
TimedPageRank	0.3398	**0.2443**	0.2514	0.2972
T-Random	0.3316	0.2054	0.2403	0.2879
T-rank	0.3356	0.2269	0.2498	0.2950
T-Fresh(1,1,30)	0.3412	0.2411	**0.2662**	**0.3076**

Table 3: Performance Comparison.

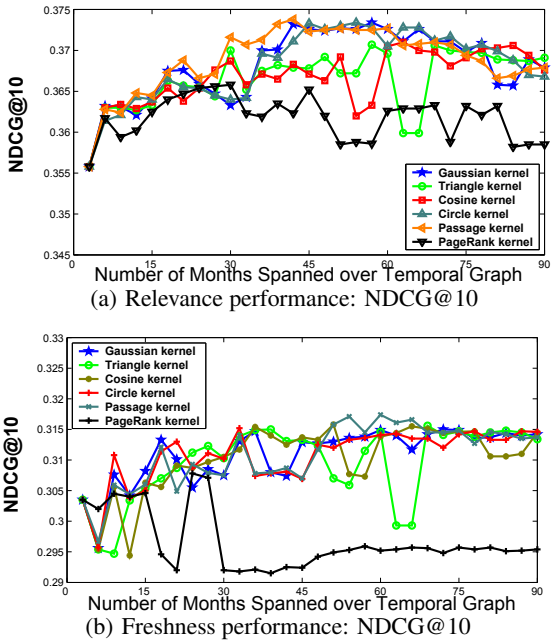
not obvious, more methods can produce fresher search results than PageRank. One reason is that these temporal link-based algorithms incorporate diverse temporal factors which favor fresh web pages. T-Fresh reaches its best P@10 at 0.3412 when γ is 0.88, which is only inferior to TemporalRank with its highest P@10 at 0.3473 when γ is 0.98. PageRank has its best P@10 at 0.3325 when γ is 0.97. With individual best combination parameter γ on P@10, we compare all the ranking algorithms over other metrics in Table 3. T-Fresh outperforms PageRank in terms of NDCG@3, NDCG@5 and NDCG@10 over 23.8%, 13.5% and 8.3%, with p-values 0.0090, 0.0260 and 0.0263 respectively. One observation is the performance of PageRank on metric NDCG@3 is extremely low while its performance on NDCG@5 and NDCG@10 are not so bad. We infer that stale web pages can achieve high authority scores by PageRank, and so dominate top positions in search results.

5.3 Deeper Analysis

We study the effects of propagation kernels and window sizes used in staying time estimation on ranking performance in this section.

Figures 4(a) and (b) show the best ranking performance of T-Fresh(*,1,*) on metric NDCG@10 for relevance and freshness. For most kernels, the relevance performance improves with the time span of the temporal graph, and reaches the highest in $[30, 60]$, i.e., from 2.5 to 5 years. The improvements upon using single snapshot are 4.9%, 4.1%, 4.2%, 4.9%, 5.0% and 2.8% for gaussian, triangle, cosine, circle, passage and PageRank kernels respectively. Passage kernel renders both a stable and best overall performance, followed by gaussian and circle kernels. Results from triangle and cosine kernels show larger fluctuations over time span of the temporal graph. Combining with the kernel expressions defined in Equations 12-17, we conclude that the ranking performance with respect to relevance can take advantage of appropriate emphasis on authority propagation among far away snapshots.

The ranking performance on freshness shows similar trends to relevance, though the variance is typically larger. Except for PageRank kernel, all others achieve their highest performance in the time interval $[30, 60]$. Passage kernel gets the best performance 0.3171 on metric NDCG@10 by outperforming the baseline (using a single snapshot) by 4.5%. One observation is that the performance of PageRank kernel suddenly falls down to around 0.295



(a) Relevance performance: NDCG@10

(b) Freshness performance: NDCG@10

Figure 4: T-Fresh(*,1,*): Sensitivity of NDCG@10 with respect to kernel for authority propagation.

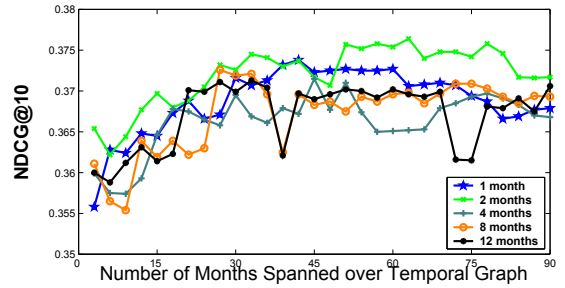
when the graph time span is beyond 30 months. One possible reason is that the authority propagation among any distinct web snapshots become very weak in PageRank kernel when the graph time span is large enough, and so historical link structures only have tiny influence on page authority estimation at the current time point. In addition, the freshness performance tends to stablize when the graph time span is over 70 months, which indicates temporal web graphs with long time span render more stable ranking performance on freshness, and it reflects the long-term freshness of web resources.

Figures 5(a) and (b) show the best ranking performance of T-Fresh(5,*,*) on metric NDCG@10 in terms of relevance and freshness. For relevance evaluation, our results demonstrate: (1) To use the average in-link freshness on several adjacent time points is better than to use it at a single time point when estimating staying time. We infer that average in-link freshness can render a good estimation about how active the page in-links are during a time period; (2) It does harm to ranking performance on relevance when the window size is too large; (3) Large window sizes result in large variance of ranking performance when varying the number of snapshots in the temporal web graph; (4) The ranking performance improves with the increase of graph time span in general for all the window sizes. For freshness evaluation, a clear trend in Figure 5(b) shows that a larger window size used in staying time estimation helps generate fresher search results with smaller deviation.

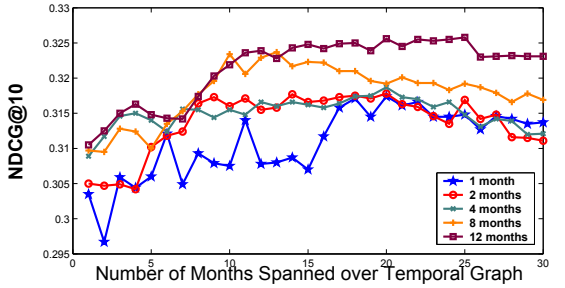
6 Conclusion and Future Work

Dynamic web resources can reflect how active web pages are over time. From the perspectives of in-links and the page itself, we quantify web freshness from web creators' activities. We argue that web freshness is an important attribute of web resources and can benefit a series of time-sensitive applications including archival search, news ranking, twitter message recommendation, tag recommendation and so on.

In this work we propose a temporal web link-based ranking algorithm to estimate time-dependent web page authority. It incorpo-



(a) Relevance performance: NDCG@10



(b) Freshness performance: NDCG@10

Figure 5: T-Fresh(5,*,*): Sensitivity of NDCG@10 with respect to window size used in the stay time estimation.

rates web freshness at multiple time points to bias the web surfer's behavior on a temporal graph composed of multiple web snapshots. Experiments on a real-world archival corpus demonstrate its superiority over PageRank on both relevance and freshness by 17.8% and 13.5% in terms of NDCG@5. Results show ranking performance can benefit more from long-term historical web freshness and link structure. The best period covers the past 2.5 to 5 years.

There are a few interesting extensions. The web surfer's temporal interest changes with her position in this work. We could fix her temporal interest, and use the estimated authority score to support archival search, which can select highly authoritative page instances that best match a specific temporal interest, in addition to relevance to a given query. On the other hand, the historical information at the Internet Archive comes from an external source for commercial search engines, which means a large amount of pages may lack archival copies. In the future, we hope to find a way to mitigate the gap caused by some pages having archival copies and some without in the searching process, so that the method can be applicable to search engines seamlessly.

Acknowledgments

We thank the anonymous reviewers for their useful comments. We also thank Fernando Diaz for valuable suggestions related to recency ranking evaluation. 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.

7 References

[1] A. Acharya, M. Cutts, J. Dean, P. Haahr, M. Henzinger, U. Hoelzle, S. Lawrence, K. Pfleger, O. Sercinoglu, and S. Tong. Information retrieval based on historical data. US Patent 7,346,839, USPTO, Mar. 2008.

[2] E. Adar, J. Teevan, S. Dumais, and J. L. Elsas. The web changes everything: understanding the dynamics of web content. In *Proc. of 2nd ACM WSDM Conf.*, pages 282–291. Feb, 2009.

[3] Amazon, Inc. Amazon mechanical turk home page, 2010. http://www.mturk.com/.

[4] E. Amitay, D. Carmel, M. Herscovici, R. Lempel, and A. Soffer. Trend detection through temporal link analysis. *Journal of the American Society for Information Science and Technology*, 55(14):1270–1281, 2004.

[5] Z. Bar-Yossef, A. Z. Broder, R. Kumar, and A. Tomkins. Sic transit gloria telae: Towards an understading of the web's decay. In *Proc. of 13th Int'l World Wide Web Conference*, pages 328–337. May 2004.

[6] K. Berberich, S. Bedathur, M. Vazirgiannis, and G. Weikum. Buzzrank... and the trend is your friend. In *Proc. of 15th Int'l World Wide Web Conference*, pages 937–938, May 2006.

[7] K. Berberich, S. Bedathur, G. Weikum, and M. Vazirgiannis. Comparing apples and oranges: Normalized PageRank for evolving graphs. In *Proc. of 16th Int'l World Wide Web Conference*, pages 1145–1146, May 2007.

[8] K. Berberich, M. Vazirgiannis, and G. Weikum. Time-aware authority ranking. *Internet Mathematics*, 2(3):301–332, 2005.

[9] J. Bian, Y. Liu, D. Zhou, E. Agichtein, and H. Zha. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *Proc. of 18th Int'l World Wide Web Conference*, pages 51–60, Apr. 2009.

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

[11] D. Cai, X. He, J. R. Wen, and W. Y. Ma. Block-level link analysis. In *Proc. of 27th Annual Int'l ACM SIGIR Conf.*, Sheffield, UK, July 2004.

[12] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1):2, 2006.

[13] J. Cho, S. Roy, and R. E. Adams. Page quality: In search of an unbiased web ranking. In *Proc. of ACM SIGMOD*, Baltimore, MD, June 2005.

[14] Y. J. Chung, M. Toyoda, and M. Kitsuregawa. A study of link farm distribution and evolution using a time series of web snapshots. In *Proc. of the 5th Int'l Workshop on Adversarial Information Retrieval on the Web*, pages 9–16, New York, NY, USA, 2009. ACM.

[15] O. de Kretser and A. Moffat. Effective document presentation with a locality-based similarity heuristic. In *Proc. of 22nd Annual Int'l ACM SIGIR Conf.*, pages 113–120, New York, NY, USA, 1999. ACM.

[16] J. L. Elsas and S. T. Dumais. Leveraging temporal dynamics of document content in relevance ranking. In *Proc. of 3nd ACM WSDM Conf.*, pages 1–10. Feb, 2010.

[17] B. Gao, T. Y. Liu, Z. Ma, T. Wang, and H. Li. A general markov framework for page importance computation. In *Proc. of 18th ACM CIKM Conf.*, pages 1835–1838, New York, NY, USA, 2009. ACM.

[18] Google Inc. Google trends home page, 2010. http://www.google.com/trends.

[19] T. Haveliwala, S. Kamvar, A. Kamvar, and G. Jeh. An analytical comparison of approaches to personalizing pagerank. Technical report, Stanford University, 2003.

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

[21] K. Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In *Proc. of 23rd Annual Int'l ACM SIGIR Conf.*, pages 41–48, July 2000.

[22] K. Kise, M. Junker, A. Dengel, and K. Matsumoto. Passage Retrieval Based on Density Distributions of Terms and Its Applications to Document Retrieval and Question Answering. In *volume 2956 of LNCS*, pages 306–327. Springer, Berlin/Heidelberg, 2004.

[23] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proc. of the ACM-SIAM Symposium on Discrete Algorithms (SODA-98)*, pages 668–677, San Francisco, CA, Jan. 1998.

[24] Y. Li and J. Tang. Expertise search in a time-varying social network. In *Proc. of 9th Int'l Web-Age Information Management Conf. (WAIM 08)*, July 2008.

[25] Y. Liu, B. Gao, T. Y. Liu, Y. Zhang, Z. Ma, S. He, and H. Li. Browserank: letting web users vote for page importance. In *Proc. of 31st Annual Int'l ACM SIGIR Conf.*, pages 451–458, New York, NY, USA, 2008. ACM.

[26] Y. Lv and C. Zhai. Positional language models for information retrieval. In *Proc. of 32nd Annual Int'l ACM SIGIR Conf.*, pages 299–306, New York, NY, USA, 2009. ACM.

[27] NIST. Text REtrieval Conference (TREC) home page, 2010. http://trec.nist.gov/.

[28] D. Petkova and W. B. Croft. Proximity-based document representation for named entity retrieval. In *Proc. of 16th ACM CIKM Conf.*, pages 731–740, New York, NY, USA, 2007. ACM.

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

[30] S. M. Ross. Introduction to Probability Models, Ninth Edition. Academic Press, Inc., Orlando, FL, USA, 2006.

[31] G. Shen, B. Gao, T.Y. Liu, G. Feng, S. Song, and H. Li. Detecting link spam using temporal information. In *Proc. of IEEE International Conference on Data Mining*, pages 1049–1053, 2006.

[32] B. Wu, V. Goel and B. D. Davison. Propagating Trust and Distrust to Demote Web Spam. In *Proc. of WWW2006 MTW Workshop*, 2006.

[33] Yahoo!, Inc. Yahoo! site explorer, 2010. http://siteexplorer.search.yahoo.com/.

[34] L. Yang, L. Qi, Y. P. Zhao, B. Gao, and T.Y. Liu. Link analysis using time series of web graphs. In *Proc. of 16th ACM CIKM Conf.*, pages 1011–1014, New York, NY, USA, 2007. ACM.

[35] P. S. Yu, X. Li, and B. Liu. On the temporal dimension of search. In *Proc. of 13rd Int'l World Wide Web Conference*, pages 448–449. ACM Press, May 2004.