# Incorporating the Surfing Behavior of Web Users into PageRank

Shatlyk Ashyralyyev
Bilkent University
Ankara, Turkey
shatlyk@cs.bilkent.edu.tr

B. Barla Cambazoglu
Yahoo! Research
Barcelona, Spain
barla@yahoo-inc.com

Cevdet Aykanat
Bilkent University
Ankara, Turkey
aykanat@cs.bilkent.edu.tr

## ABSTRACT

In large-scale commercial web search engines, estimating the importance of a web page is a crucial ingredient in ranking web search results. So far, to assess the importance of web pages, two different types of feedback have been taken into account, independent of each other: the feedback obtained from the hyperlink structure among the web pages (e.g., PageRank) or the web browsing patterns of users (e.g., BrowseRank). Unfortunately, both types of feedback have certain drawbacks. While the former lacks the user preferences and is vulnerable to malicious intent, the latter suffers from sparsity and hence low web coverage. In this work, we combine these two types of feedback under a hybrid page ranking model in order to alleviate the above-mentioned drawbacks. Our empirical results indicate that the proposed model leads to better estimation of page importance according to an evaluation metric that relies on user click feedback obtained from web search query logs. We conduct all of our experiments in a realistic setting, using a very large scale web page collection (around 6.5 billion web pages) and web browsing data (around two billion web page visits).

## Categories and Subject Descriptors

H.3.3 [**Information Storage Systems**]: Information Retrieval Systems

## General Terms

Algorithms, Performance, Experimentation, Human Factors

## Keywords

Page quality, web search, ranking, PageRank, BrowseRank

## 1. INTRODUCTION

Query-dependent features, such as BM25, are successfully used in IR systems to estimate the degree of relevance between a given query and a document. In the context of large-scale web search engines, however, quantifying only the relevance is not adequate. The large size of the Web and high variation in content quality require distinguishing the importance of web pages independent of the query. Hence, most web search engines incorporate query-independent page importance scores into their ranking algorithms.

PageRank [22] is perhaps the most well-known and widely used technique for computing web page importance. The basic idea behind this technique is to compute the importance of a web page based on the quantity of the links received from other pages as well as the quality of those referring pages. Although PageRank has found many important use cases, it has two serious drawbacks. First, PageRank solely relies on the hyperlink structure of the Web without incorporating any kind of feedback from the real users surfing the Web. Therefore, all pages are treated equally, ignoring their importance for end users or the likelihood of being visited by a web surfer [9]. Second, since the hyperlink structure is mainly created by the web site owners, it is subject to manipulation. As an example, link farms can be created to artificially boost the importance of certain web pages, making PageRank vulnerable to link spam [12].

An interesting alternative to PageRank is to exploit the web surfing behavior of users to assess the importance of web pages (e.g., BrowseRank [21]). In this approach, a virtual link structure is created between web pages based on the web browsing patterns of users, i.e., the transitions they make between different pages when surfing the Web. Such patterns can be obtained by mining navigational user activity that is tracked by the toolbar applications, commonly installed in web browsers. This approach provides better quality feedback about page importance and also solves the previously mentioned spam problem associated with PageRank. However, the web browsing patterns extracted from the toolbar logs are very sparse. Hence, many web pages (especially, the less popular web pages) are not covered and their scores cannot be computed.

One of the main objectives of this work is to investigate whether exploiting web and user feedback at the same time (i.e., using both web data and browsing data) improves the quality of page rankings over using only one type of feedback. To this end, we define a discrete-time Markov chain constructed by aggregating web and browsing data with properly scaled page transition probabilities. Importance scores of pages are estimated using the standard procedure followed in PageRank computations. We refer to the proposed technique as PBRank (PageBrowseRank). All of our experiments are conducted in a large scale and realistic setting.


*CIKM'13,* October 27 - November 1, 2013, San Francisco, CA, USA
 ACM 978-1-4503-2263-8/13/10 $15.00.

Our contributions can be summarized as follows:

- We devise a hybrid ranking model that uses a mixture of feedback obtained from the hyperlink structure of the Web and the web browsing patterns of users.
- We shed light into the overlap between the web data, browsing data, and web search click data as well as the correlation between the importance values assigned to web hosts by these data sources.
- We experiment in a realistic setting with very large data, orders of magnitude larger than the data used in earlier works in the same problem context.

The following are the selected findings of our work:

- Exploiting both web and user feedback at the same time improves the quality of the page ranking compared to using only one type of feedback.
- Using the web data increases the coverage (the number of web hosts for which an importance score can be computed) over using only the browsing data.
- When the web and user feedbacks are optimally combined, the user feedback has 99 times more influence on the quality of page rankings than the web feedback.

The rest of the paper is organized as follows. In Section 2, we provide some background on the PageRank and BrowseRank techniques. The proposed hybrid ranking model, PBRank, is presented in Section 3. We present our performance evaluation metrics in Section 4. Then, in Section 5, we provide the characteristics of our data. The results of conducted experiments are presented in Section 6. A brief survey of related work is given in Section 7. Finally, we conclude the paper in Section 8.

## 2. BACKGROUND

**Pagerank.** PageRank [22] is motivated by the academic citation literature. The main idea in this technique is to assign higher scores to pages that receive many links from other important pages which have relatively few out-links [2]. The computation of scores relies on a probabilistic model known as the random surfer model, where the score of a page is defined by the steady-state probability that the surfer will be at that particular page at some time step in the future. This model consists of a Markov chain induced by a random walk on a web graph having $n$ vertices. Each state of the chain corresponds to a different vertex in the web graph. A transition matrix $\mathbf{P}=(p_{ij})$ is associated with this chain such that

$$p_{ij}=\begin{cases} 1/|\mathcal{L}_i|, & |\mathcal{L}_i|>0; \\ 0, & \text{otherwise.} \end{cases}, \tag{1}$$

where $\mathcal{L}_i$ denotes the set of out-links of page $i$. Given this transition matrix, the PageRank vector $\mathbf{p}=(p_i)$, where $p_i$ indicates the score of page $i$, can be computed by finding the Markov chain's stationary distribution that satisfies $\mathbf{p}=\mathbf{P}^{\mathrm{T}}\mathbf{p}$, i.e., the principal eigenvector of the chain. The solution can be obtained through a series of iterations of the form $\mathbf{p}^{k+1}=\mathbf{P}^{\mathrm{T}}\mathbf{p}^{k}$ using the power method [10]. The existence of a solution, i.e., the convergence of iterations, requires the $\mathbf{P}$ matrix to be stochastic, irreducible, and aperiodic, neither of which are guaranteed for $\mathbf{P}$.

The reason behind matrix $\mathbf{P}$ for not being stochastic is the presence of sink (or so-called dangling) pages with no out-links. Although there are other possibilities [3, 16, 17], the common solution to this problem is to add artificial links from such pages to every other page in the Web [22]. This results in a stochastic transition matrix $\mathbf{P}'$, computed as

$$\mathbf{P}'=\mathbf{P}+\mathbf{d}\mathbf{v}^{\mathrm{T}}, \tag{2}$$

where $\mathbf{d}=(d_i)$ is a dangling page vector (if $i$ is a dangling page, $d_i=1$; otherwise, $d_i=0$) and $\mathbf{v}=(v_i)$ is a vector, where $v_i$ indicates the transition probability from dangling pages to a specific page $i$. Typically, the transition probabilities are set to $1/n$ for all pages. The resulting matrix $\mathbf{P}'$ is stochastic, but not irreducible. Applying a similar technique on $\mathbf{P}'$, an irreducible stochastic transition matrix $\mathbf{P}''$ can be obtained, also guaranteeing aperiodicity as

$$\mathbf{P}''=\alpha\mathbf{P}'+(1-\alpha)\mathbf{e}_n\mathbf{t}^{\mathrm{T}}. \tag{3}$$

Here, $\mathbf{e}_n$ is a vector of size $n$ containing all ones. $\alpha$ denotes the probability that the surfer will follow one of the links in the current page while (1–$\alpha$) is the probability that the surfer will jump to a page that is not necessarily linked by the current page. In practice, $\alpha$ values between 0.85 and 0.9 are used. The $\mathbf{t}=(t_i)$ vector is referred to as the teleportation vector, where $t_i$ indicates the probability of jumping to page $i$. Typically, this probability is set to $1/n$ for all pages. In case of personalized or topical teleportation vectors, non-uniform jump probabilities can also be used [13].

**BrowseRank.** BrowseRank [21] mainly relies on the same principles with PageRank. However, it is based on a continuous-time Markov process, which exploits the staying times of users on pages. In this technique, the browsing graph is constructed as $\mathcal{G}=(\mathcal{V},\mathcal{W})$, where $\mathcal{V}$ is the set of vertices representing web pages visited by users and $\mathcal{W}$ is a set of directed edges indicating the visit patterns between pages. Vertices are associated with staying times of users on respective pages and the teleportation probabilities of those pages. Each edge is weighted with the number of visits between the two pages corresponding to its end vertices.

Given the above-mentioned browsing graph, a continuous-time Markov process is defined. This model is then converted into a discrete-time Markov process whose stationary distribution is estimated using the power method. The details of this conversion are too technical to be discussed here and we refer the reader to [21] for further details. Instead, herein, we briefly discuss how the transitions of users between different pages are obtained since this step is somewhat different than the procedure we adopted in our work.

In BrowseRank, the transitions that users make between different pages are obtained by relying on the page visits observed in individual web browsing sessions of users. A browsing session is composed of a series of page visits that are sorted in increasing order of timestamps. Initially, an edge $(v_i, v_j)$ is added to $\mathcal{W}$ for every pair $(i,\ j)$ of pages such that there is no other page visited between $i$ and $j$. Afterwards, an edge $(v_i, v_j)$ is removed from $\mathcal{W}$ if the user visited page $j$ by typing its URL in the browser's navigation bar or if page $j$ is visited more than 30 minutes after page $i$ is visited.

## 3. PBRANK

The main idea behind PBRank is to combine two different types of feedback, i.e., those provided by the web data and browsing data in a meaningful way. Our goal is to come up with a simple extension to the standard procedure summarized in Section 2, leaving the theoretical foundations unchanged. To this end, we use a transition matrix $\mathbf{X}$ corresponding to the pages in the union of the web and browsing

data. $\mathbf{X}$ is a square matrix of size $m \times m$ and is expressed as a linear combination of two other matrices of the same size:

$$\mathbf{X} = \lambda \mathbf{P}'' + (1 - \lambda)\mathbf{B}''. \quad (4)$$

Here, $\mathbf{P}''$ is an $m \times m$ version of the final PageRank matrix used in the power method iterations (see Eq. 3), i.e., this matrix is created based on the web feedback. In addition, using the user feedback, we define another matrix $\mathbf{B}''$, which we will describe next. $\lambda$ is a constant in the $[0, 1]$ range and is used to adjust the influence of one type of feedback over the other. The page importance scores can be obtained by finding the principal eigenvector of $\mathbf{X}$ using the power method as usual.

We form the $\mathbf{B}''$ matrix in a similar fashion to Eq. 3:

$$\mathbf{B}'' = \beta \mathbf{B}' + (1 - \beta)\mathbf{e}_n \mathbf{r}^{\mathrm{T}}, \quad (5)$$

where $\beta$ and $\mathbf{r} = (r_i)$ are the counterparts of the $\alpha$ constant and the $\mathbf{t}$ vector in Eq. 3, respectively. We use biased teleportation probabilities in $\mathbf{r}$, instead of uniformly setting them to $1/n$ as in $\mathbf{t}$. The teleportation probability $r_i$ of a particular page $i$ is computed as

$$r_i = \frac{1 + T_i}{m + \sum_{j=1}^{m} T_j}, \quad (6)$$

where $T_i$ denotes the number of visits to page $i$ by means other than following a link in a page. This way, the jumping behavior of the surfer is biased towards more popular pages. Here, we add one to visit counts for smoothing purposes.

Following the idea in [9], $\beta$ can be computed as

$$\beta = \frac{\sum_{j=1}^{m} (V_j - T_j)}{\sum_{j=1}^{m} V_j}, \quad (7)$$

where $V_j$ denotes the total visit count of page $j$. The $\beta$ constant reflects the users' tendency to reach a page by following the hyperlinks in web pages.

The $\mathbf{B}'$ matrix is computed by the following equation:

$$\mathbf{B}' = \mathbf{B} + \mathbf{d}\mathbf{v}^{\mathrm{T}}, \quad (8)$$

where $\mathbf{d}$ and $\mathbf{v}$ are defined as before (see Eq. 2). The probabilities in the page transition matrix $\mathbf{B} = (b_{ij})$ are set depending on the likelihood of a hyperlink being followed by users. Therefore, the links within a page are not treated equally as in Eq. 1. Instead, the transition probability from page $i$ to page $j$ is computed in a biased manner by taking into account the share of the click volume of page $j$ in the overall click volume observed on page $i$ as

$$b_{ij} = \frac{V_{ij}}{\sum_{k \in \mathcal{L}_i} V_{ik}}, \quad (9)$$

where $V_{ij}$ is the click volume from page $i$ towards page $j$.

PBRank can be considered as a variant of BrowseRank since both techniques use page visit probabilities extracted from browsing data. In practice, one may prefer PBRank to BrowseRank because of the following reasons. First, as we will show later in Section 6, PBRank achieves a better coverage of web pages than BrowseRank due to the use of web data in scoring computations, i.e., a larger number of pages receive non-zero scores. Second, PBRank's implementation is easier than the implementation of BrowseRank, which employs a relatively more sophisticated continuous-time Markov model. Finally, the transition probabilities computed in PBRank are accurate values computed over actual user clicks on links. The transition probabilities computed in BrowseRank, however, are only approximations because they are computed based on a timestamp-sorted sequence of page visits in user sessions, not the links that are actually followed by users. Given that many users browse the Web by opening multiple browser tabs [14] and concurrently following links in different tabs, a time-ordered sequence of page visits may not be sufficient to obtain the actual transitions between pages. Hence, the transition probabilities computed in BrowseRank may not reflect the true surfing patterns of users.

We note that the existence of a solution is guaranteed since the $\mathbf{X}$ matrix is irreducible and aperiodic because both summation terms in Eq. 4 already have these properties. When $\lambda = 0$ or $\lambda = 1$, $\mathbf{X}$ may not be row-stochastic, but this does not prevent the convergence of iterations. If $\lambda$ is set to zero or one in Eq. 4, PBRank reduces to a discrete-time variant of BrowseRank or PageRank, respectively. As we will see in Section 6, the best ranking quality will be obtained for $\lambda$ values close to zero.

# 4. EVALUATION METRICS

Given different ranking techniques for estimating page importance, we would like to quantify two different aspects of those techniques: page coverage and ranking quality. The former aspect refers to the ability of the ranker to compute a (non-zero) score for many pages. The second aspect refers to the ability of the technique to place important pages at higher ranks with respect to a ground-truth ranking.

Herein, we define two separate metrics to quantify the two aspects mentioned above. In both metrics, to represent the actual page importance, we rely on a ground-truth data obtained from web search click logs. Since our focus in this work is on the impact of the generated page rankings on web search, the user clicks issued on web search results form a natural ground-truth. We note that, at the level of individual queries, the click likelihood of a page in search results is largely affected by the relevance of the page to the query. However, as the click information is aggregated over many different queries, the click volume of a page in search results stands as a fairly reasonable ground-truth indicating the importance of the page for users.

We now introduce some notation and define our metrics. Let $\rho$ be a given page ranking technique and $\mathcal{R}^{\rho}$ be the set of pages that are accessible by this technique. Every page in $\mathcal{R}^{\rho}$ when ranked by $\rho$ receives a positive importance score, higher scores indicating more important pages. Let $\rho^*$ be an oracle ranker that can access to the ground-truth importance values for a set $\mathcal{R}^*$ of pages. Given these definitions, the **coverage** $\chi^{\rho}$ of a ranking technique $\rho$ is defined as the fraction of ground-truth pages for which a score can be computed by $\rho$.

Next, we devise a metric to quantify the quality of the ranking generated by a given ranker $\rho$ with respect to the oracle ranker $\rho^*$. In literature, there are various metrics readily available for measuring the correlation between two given rankings (e.g., Kendall's tau [18] and Spearman's footrule [24]). Unfortunately, those metrics operate on fully ranked lists. In our case, we have partially ranked lists (i.e., some pages in the ground-truth set are not available to ranker $\rho$ and vice versa). In literature, there are variants that can handle partial rankings [7]. Nevertheless, we prefer

not to use those metrics here because neither of them take into account the popularity of the ranked items (i.e., in our setting, the click volumes of pages would be omitted). In our setting, the commonly used IR metrics such as DCG or NDCG [15] are not very useful either because, in such metrics, the top ranked items are heavily weighted. Our rankings are very long and having such a strong bias only at the top ranks is not very meaningful.

Due to the above-mentioned reasons, we devise a quality metric that can capture items' estimated rank and the importance in the ground-truth data at the same time while being able to yield meaningful results for rankings with a large number of items. We define the relative quality $\Phi^{\rho}(k)$ of a given ranking $\mathcal{R}^{\rho}$ at rank $k$ with respect to a ground-truth ranking $\mathcal{R}^{*}$ as

$$\Phi^{\rho}(k) = \frac{\phi^{\mathcal{R}^{\rho}}(k)}{\phi^{\mathcal{R}^{*}}(k)}. \quad (10)$$

Here, $\phi^{\mathcal{R}^{*}}(k)$ is a normalization factor representing the quality of the best possible ranking that can be achieved by the oracle ranker $\rho^{*}$. Note that the best possible ranking is achieved when pages in $\mathcal{R}^{*}$ are ranked in decreasing order of their importance. We define $\mathcal{C}^{\mathcal{R}}(k)$ (the sum of importances of top $k$ pages in $\mathcal{R}$) and $\phi^{\rho}$ through recursive functions as

$$\mathcal{C}^{\mathcal{R}}(k) = \begin{cases} 0, & \text{if } k = 0; \\ \mathcal{C}^{\mathcal{R}}(k-1) + I(\mathcal{R}_k), & \text{if } 1 \leq k \leq |\mathcal{R}|; \\ \mathcal{C}^{\mathcal{R}}(|\mathcal{R}|), & \text{if } k > |\mathcal{R}|. \end{cases}, \quad (11)$$

$$\phi^{\mathcal{R}}(k) = \begin{cases} 0, & \text{if } k = 0; \\ \phi^{\mathcal{R}}(k-1) + & \\ \mathcal{C}^{\mathcal{R}}(k-1) + \frac{I(\mathcal{R}_k)}{2}, & \text{if } 1 \leq k \leq |\mathcal{R}|; \\ \phi^{\mathcal{R}}(k-1) + & \\ \mathcal{C}^{\mathcal{R}}(k-1), & \text{if } k > |\mathcal{R}|. \end{cases}, \quad (12)$$

where $\mathcal{R}_k$ denotes the $k$-th ranked page in a given ranking $\mathcal{R}$ of pages and $I(p)$ denotes the importance of page $p$, inferred from the ground-truth data. In Section 6, we will consider two different alternatives to compute $I(p)$, one assuming unit page importance values and another where the page importance is represented by the click volume of the page in search results. In either case, we will assume that $I(p) = 0$ if $p \notin \mathcal{R}^{*}$. We note that PageRank, BrowseRank, and PBRank yield only positive scores when ranking pages.

The devised $\phi$ metric emphasizes accumulation of importance at early ranks, as the total page importance attained at rank $k$ continues to contribute to the value of the metric at all ranks following $k$. In this sense, the functioning of this metric resembles the ROC analysis and the area under the curve metric [8].

# 5. DATA

**Web page collection.** We use a snapshot of the Web (crawled in late 2011), which contains around 6.5 billion web pages. Due to the difficulties involved in parsing web pages written in the CJK languages, we exclude such pages from further consideration. Moreover, self-links are removed and identical out-links in a page are contracted into a single out-link. We convert the remaining pages and links into a web graph and further compress this graph to obtain a host-level graph of the Web. In the rest of the paper, we use this host-level graph, which includes about 230 million unique web hosts with 1.5 billion inter-host links.

**Browsing data.** We obtain the web browsing data through a commercial toolbar application deployed at the web scale on a large number of web browsers. Our experiments use only the browsing data acquired from users who explicitly gave permission for their page views to be logged. For each visited page, the toolbar log contains information about the time at which the page is visited and how the page is reached. In particular, a page has a referrer URL if the user reached the page by clicking a link in another page, otherwise, i.e., if the user manually typed the URL or clicked a bookmark link, the referrer URL is not available. In total, our data contains around two billion page visits, performed by users all around the world. The browsing data is obtained in a period right after the web collection is obtained.

**Click data.** To serve as a ground-truth in evaluation of the generated page rankings, we use a large (random) sample of clicks obtained from the query logs of a commercial web search engine. The query log contains information about the query string, and the URLs clicked by the user who submitted the query. The data includes around 170,000 unique URLs and over 700,000 clicks. The click sample is obtained in a time period that follows the acquisition of the browsing data. The scatter plot in Fig. 1 shows the correlation between the visit counts of URLs in the browsing data and their click counts. According to this figure, there is a large number of URLs that are highly visited by web surfers, but not received many clicks from search engine users when displayed in web search results. However, the reverse statement is not true, i.e., highly clicked web pages tend to be visited by many web surfers. This observation provides us enough motivation to use the click data as an alternative ground-truth for representing page importance.

# 6. EXPERIMENTS

**Optimizing $\beta$.** An important parameter that needs to be tuned before PBRank computations is the $\beta$ constant. Tuning this constant requires measuring the ratio between the number of visits initiated by following an out-link in a page and the total number of visits in the browsing data (see Eq. 7). In our data, we found this value to be $\beta = 0.62$, i.e., URLs are slightly more likely to be visited by following links. The obtained number is somewhat consistent with the earlier observation in [9]. Fig. 2 displays the distribution of URL visit counts in the browsing data. We observe a power-law distribution with slight distortion at the head.

**Overlap among data sources and Coverage.** Fig. 3 displays the overlap between the three different data sources. As expected, the web data is considerably larger than the browsing and click data. Based on these numbers, we can compute the coverage metric ($\chi$), defined in Section 4. While using only the browsing data provides a coverage of 80.1%, using only the web data gives a coverage of 95.8%. On the other hand, using both of the feedbacks results in a coverage of 97.2% and provides a coverage increase of 17.1% over using only the browsing data. This result indicates that PBRank can produce non-zero importance scores for a larger number of URLs than both BrowseRank and PageRank.

**Optimizing $\lambda$.** We next aim to find the $\lambda$ value that optimizes the ranking quality metric defined in Eq. 12. To this end, we compute the value of this metric for different PBRank rankings that are obtained by varying $\lambda$ through parameter sweeping. Fig. 4 shows the values of the metric with $\lambda$ increased between zero and one at increments of 0.1.

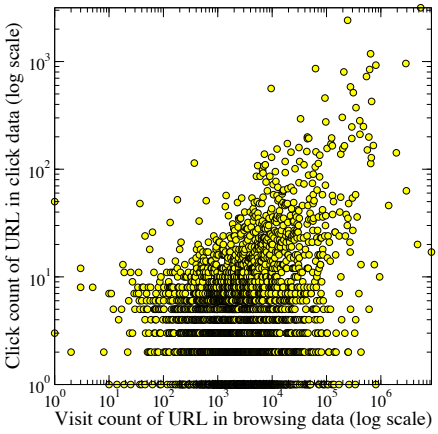


**Figure 1: Visit count of a URL in the browsing data versus its click count in search results.**

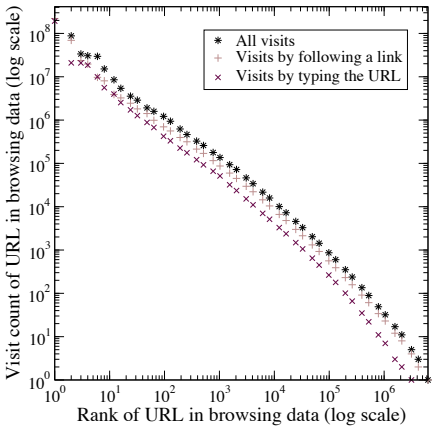


**Figure 2: Distribution of URL visit counts in toolbar data.**

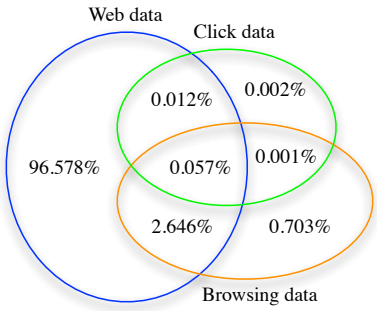


**Figure 3: Distribution of all available URLs in the web data, browsing data, and click data.**

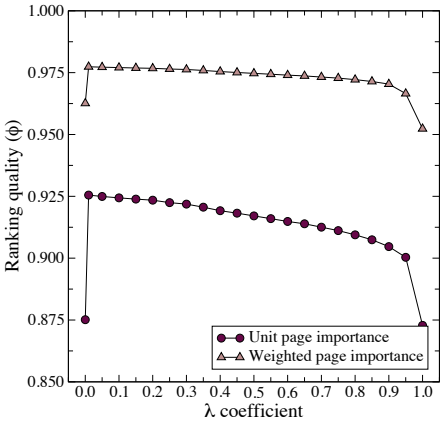


**Figure 4: The variation in ranking quality ($\Phi$) for different values of $\lambda$.**

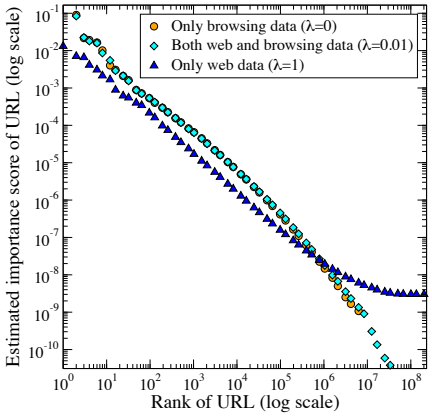


**Figure 5: Distribution of URL importance scores.**

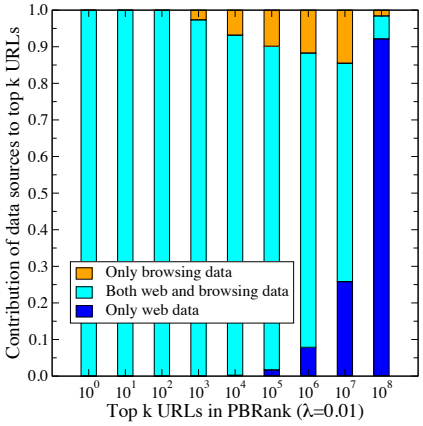


**Figure 6: Contribution of different data sources to the top $k$ URLs in PBRank with $\lambda = 0.01$.**

According to the figure, any $\lambda$ value between zero and one yields a superior ranking performance than either baseline. We observe better performance as $\lambda$ is closer to zero. Hence, we perform another parameter sweep for $\lambda$ values near zero. The results of this experiment are displayed in Table 1. We observe that the optimum $\lambda$ value is somewhere between 0.005 and 0.015. In the rest of the experiments, we set $\lambda$ to 0.01, where we observe the best ranking quality (when the page importance is weighted by the click count). According to the ratio 0.99/0.01, the feedback obtained from the browsing data has 99 times more influence on the ranking quality than the feedback coming from the web data.

**Comparison of rankings.** Fig. 5 shows the distribution of URL importance scores generated by PBRank for three different values of $\lambda$. All distributions are heavily skewed. As expected, the score distributions for $\lambda = 0$ and $\lambda = 0.01$ are very similar to each other and somewhat different than the score distribution in case of $\lambda = 1$. The curve representing $\lambda = 0$ is shorter than the other two because fewer URLs (only those in the web browsing data) are ranked.

**Contribution of data sources to top $k$ ranks.** As illustrated in Fig. 6, the main contribution to the top $k = 100$ URLs in the ranking generated by PBRank ($\lambda = 0.01$) comes from the URLs that are present in both web and browsing data. As $k$ increases to 10K, we observe some URLs that are available only in the browsing data to enter the rankings. The URLs that are available only in the web data become visible after the first 10K ranks. This result indicates that the URLs in the very top ranks are mainly determined by the feedback obtained from the browsing data.

## 7. RELATED WORK

**Overview.** PageRank is originally proposed in [22]. The technique finds application in a variety of problems from different domains including bibliometrics [6], web crawling [5], and spam detection [11]. HITS [19] is a technique closely related to PageRank. Interested reader may refer to [2] and [20] for a survey of further issues.

**Customizing PageRank.** A large effort is spent to customize PageRank computations depending on the interests of users. This is mainly achieved by either adjusting the $\alpha$ constant, which shows the probability of following a link in the current page, or by customizing the teleportation vector $\mathbf{v}$ (see Eq. 3). Regarding the first possibility (customizing the random jump probability), several works investigated the effect of $\alpha$ on the quality of the final rankings [1, 4,

**Table 1: The ranking quality metric ($\Phi$) for varying values of $\lambda$**

| | $\Phi$ | |
|---|---|---|
| $\lambda$ | Unit weight | Weighted |
| (only browsing data) 0 | 0.87512 | 0.96259 |
| 0.00001 | 0.92265 | 0.97637 |
| 0.0001 | 0.92390 | 0.97679 |
| 0.001 | 0.92496 | 0.97716 |
| 0.005 | 0.92536 | 0.97731 |
| 0.01 | **0.92550** | **0.97738** |
| 0.015 | 0.92541 | 0.97735 |
| 0.02 | 0.92531 | 0.97733 |
| 0.03 | 0.92525 | 0.97730 |
| 0.04 | 0.92504 | 0.97725 |
| 0.05 | 0.92494 | 0.97723 |
| 0.1 | 0.92437 | 0.97705 |
| (only web data) 1 | 0.87283 | 0.95232 |

9, 23]. The order of pages in the final PageRank vector is found to be heavily affected by the $\alpha$ constant used [23]. The results reported in [4] show that $\alpha$ values close to 1 do not yield accurate rankings. Two latter works suggest using $\alpha$ values around 0.5 [1] or in the 0.6–0.725 range [9]. The approach proposed in [9] is relevant to ours in that it relies on the web browsing data to set the $\alpha$ constant. Regarding the second possibility (customizing the teleportation vector), several attempts were made [13, 16, 17]. In topic-sensitive PageRank [13], several topic-specific PageRank vectors are computed for a fixed number of topics. The PageRank computation is biased to yield higher scores for pages belonging to a certain topic by simply adjusting the jump probabilities in the teleportation vector. In [17], a similar idea is described, restricting personalization preferences to blocks of web domains instead of topics. In [16], an approximate personalized PageRank vector is computed based on precomputed basis vectors. The BrowseRank approach [21] relies on web browsing data to customize the teleportation vector.

Our work goes beyond these works in different aspects. First, we use web browsing data of users to customize the probabilities in the transition matrix, instead of adapting only the $\alpha$ constant as in [9] or adjusting the probabilities in the teleportation vector as in [21]. In this respect, our model can accurately capture the variation in the quality of the links within web pages, unlike the above-mentioned two works, which assume a uniform probability for following a link in a page. Moreover, we conduct our experiments in a very large setting, orders of magnitude larger than the settings in most previous work.

# 8. CONCLUSIONS

We proposed a novel model for computing web page importance scores by using a mixture of the feedback extracted from the hyperlink structure of the Web and the feedback obtained from the web browsing patterns of users. According to a quality metric using user clicks on web search results mined from a query log, the proposed hybrid model exploiting both the web structure and the navigation patterns of users lead to a better performance than using only a single type of feedback. We found that the optimum mixture is achieved when 99% of the score comes from the browsing feedback, and only 1% from the web feedback.

# 9. REFERENCES


[1] K. Avrachenkov, N. Litvak, and K. S. Pham. Distribution of PageRank mass among principle components of the Web. In *Proc. 5th Int'l Conf. Algorithms and Models for the Web-Graph*, pages 16–28, 2007.

[2] P. Berkhin. A survey on PageRank computing. *Internet Mathematics*, 2(1):73–120, 7 2005.

[3] M. Bianchini, M. Gori, and F. Scarselli. Inside PageRank. *ACM Transactions on Internet Technology*, 5(1), 2005.

[4] P. Boldi, M. Santini, and S. Vigna. PageRank as a function of the damping factor. In *Proc. 14th Int'l Conf. World Wide Web*, pages 557–566, 2005.

[5] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through URL ordering. In *Proc. 7th Int'l Conference on World Wide Web*, pages 161–172, 1998.

[6] Y. Ding, E. Yan, A. Frazho, and J. Caverlee. PageRank for ranking authors in co-citation networks. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2229–2243, 2009.

[7] R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing partial rankings. *SIAM Journal on Discrete Mathematics*, 20(3):628–648, 2006.

[8] T. Fawcett. An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27(8):861–874, 2006.

[9] D. F. Gleich, P. G. Constantine, A. D. Flaxman, and A. Gunawardana. Tracking the random surfer: empirically measured teleportation parameters in PageRank. In *Proc. 19th Int'l Conf. World Wide Web*, pages 381–390, 2010.

[10] G. H. Golub and J. F. V. Loan. *Matrix Computation*. John Hopkins University Press, 3 edition, 1996.

[11] Z. Gyongyi, P. Berkhin, H. Garcia-Molina, and J. Pedersen. Link spam detection based on mass estimation. In *Proc. 32nd Int'l Conf. Very Large Data Bases*, pages 439–450, 2006.

[12] Z. Gyöngyi and H. Garcia-Molina. Link spam alliances. In *Proc. 31st Int'l Conf. Very Large Data Bases*, pages 517–528, 2005.

[13] T. H. Haveliwala. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. *IEEE Trans. Knowl. Data Eng.*, 15(4):784–796, 2003.

[14] J. Huang and R. W. White. Parallel browsing behavior on the Web. In *Proc. 21st ACM Conf. Hypertext and Hypermedia*, pages 13–18, 2010.

[15] K. Jarvelin and J. Kekalainen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[16] G. Jeh and J. Widom. Scaling personalized web search. In *Proc. 12th Int'l Conf. World Wide Web*, pages 271–279, 2003.

[17] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Exploiting the block structure of the Web for computing PageRank. Technical report, Stanford University, 2003.

[18] M. G. Kendal. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[19] J. Kleinberg. Authoritive sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[20] A. Langville and C. Meyer. Deeper inside PageRank. *Internet Mathematics*, 1(3):335–380, 2005.

[21] 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. 31st Annual Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pages 451–458, 2008.

[22] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the Web. Technical report, Stanford University, 1998.

[23] L. Pretto. A theoretical analysis of Google's PageRank. In *Proc. 9th Int'l Symp. String Processing and Information Retrieval*, pages 131–144, 2002.

[24] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.