Mining Rich Session Context to Improve Web Search

Guangyu Zhu[*]
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
zhugy@umiacs.umd.edu

Gilad Mishne
Yahoo! Labs
701 First Ave.
Sunnyvale, CA 94089
gilad@yahoo-inc.com

ABSTRACT

User browsing information, particularly their non-search related activity, reveals important contextual information on the preferences and the intent of web users. In this paper, we expand the use of browsing information for web search ranking and other applications, with an emphasis on analyzing individual user sessions for creating aggregate models. In this context, we introduce *ClickRank*, an efficient, scalable algorithm for estimating web page and web site importance from browsing information. We lay out the theoretical foundation of ClickRank based on an intentional surfer model and analyze its properties. We evaluate its effectiveness for the problem of web search ranking, showing that it contributes significantly to retrieval performance as a novel web search feature. We demonstrate that the results produced by ClickRank for web search ranking are highly competitive with those produced by other approaches, yet achieved at better scalability and substantially lower computational costs. Finally, we discuss novel applications of ClickRank in providing enriched user web search experience, highlighting the usefulness of our approach for non-ranking tasks.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Theory

Keywords

ClickRank, aggregate user behavior, intentional surfer model, learning to rank, web search

1. INTRODUCTION

Knowledge discovery and mining of user behavior data on the web promises major improvements in several key aspects of web search. Usage information associated with web search, such as aggregated user activities on search engine result pages, has been a valuable source of information for learning and recognizing query intents. Studies of search engine query logs [36, 35, 2, 37, 1, 25] and web search click-through data [9, 20, 21, 3, 33] have demonstrated significant improvements in retrieval quality, even when the activities examined are limited to actions on search result pages only – a relatively small fraction of a user's activity online.

By effectively incorporating information on *all* web user activities, search engines gain insights into user preferences and intents, and improve both retrieval performance and user experience. First, analysis of all user actions provides a more robust estimate of user perceived importance associated with web pages and sites [26]. We discuss this aspect in more details later on. Second, search engines face the challenge of prioritizing and adapting their computing resources under practical constraints in crawling, indexing, and query processing [4]. In this context, the relative attention a web page receives from all users provides an intuitive and user-centric optimization criterion, and is responsive to evolving user behaviors. As a large amount of web content emerges and refreshes within a shorter time interval than a typical crawling and indexing cycle of a search engine [31], discovering popular content and adapting crawling schedules based on the degree of usage may prove to be an effective and agile policy. Finally, another challenging area for search engines is access to the deep (or invisible) web – the fraction of the web that is dynamically generated and not directly accessible to automated crawlers [17]. Their coverage can substantially improve by leveraging large-scale user browsing history, which collectively reveals some of the hidden URLs, providing gateways to their content.

In this paper, we focus on using the large amount of knowledge gained from computational analysis of user browsing behavior by: (1) leveraging all browsing actions; and (2) developing models that incorporates rich context within logical units of user activity—user sessions. Our main contribution is ClickRank, a novel algorithm we propose for estimating web page and web site importance, which is based on these two key notions. ClickRank first estimates a local importance value for every page or site in each user browsing session, based on the implicit preference judgments of the user in the session context. It then aggregates these local values over all sessions of interest to create a global ranking.

We evaluate this approach in three important areas of web search. Our first experiment is on the traditional task of website ranking, where we show that results from ClickRank are competitive against state-of-the-art approaches, including PageRank [32] and the recently proposed BrowseRank [26], and yet obtained at significantly lower computational costs. In the second experiment, we demonstrate the novelty and effectiveness of ClickRank in web page ranking together with several hundred state-of-the-art web search features, including those computed from page visit counts and the link structure of the web graph. In this large-scale test, we formulate

[*]This work was conducted at Yahoo! Inc.

KDD'09, June 28–July 1, 2009, Paris, France.

the task of learning the optimal ranking model as an additive regression problem using gradient boosted decision trees, and quantify the variable importance of ClickRank in direct comparison with other features. Finally, we demonstrate ClickRank in a system that mines and presents recent, popular pages to web search users as dynamic quicklinks in search result summaries.

The structure of this paper is as follows. Section 2 reviews related work. In Section 3, we present important characteristics of general web sessions, and describe in details our approach to session mining by incorporating contextual information in session representation. In Section 4, we introduce the ClickRank algorithm, and describe how we combine it with existing approaches for the task of web search by learning an optimal web search ranking model. We comprehensively evaluate ClickRank in three core web search applications in Section 5 and conclude in Section 6.

2. RELATED WORK

PageRank [32], HITS [23], and TrustRank [16] are representative link analysis algorithms for computing authoritative sources using the link structure of the web graph, and have been widely used in web search as measures of relative importance of web pages. The well-known PageRank algorithm, for instance, considers a link from a source page to another as an explicit endorsement of the destination page in perceived page quality, and uses only the static link structure of the web as input. Based on the assumption of a random surfer model and the first-order Markov process, PageRank computes the stationary probability distribution for the web link graph iteratively, resulting in the World's largest matrix computation [29].

A number of problems are commonly associated with link analysis algorithms. First, user browsing behaviors are driven by intents, and they significantly deviate from the random surfer model that PageRank is based on. A recent study on real network traffic [28] demonstrated that user visitation patterns differs considerably from that approximated by the uniform surfing behavior model used in PageRank. Second, static modeling of the link structure favors old pages, because a new page is less likely to be linked to within a short period of time, even if it has very good quality. Third, link structures are prone to manipulation as adversarial links can be generated to artificially inflate ranking more quickly than quality links that typically originate in manual editing. Last, as the web grows at an explosive speed[1], computing page importance at the web scale by link analysis becomes very computationally expensive [24], even through various optimization schemes [8, 27].

As a logical unit of general user web experience, a web session contains rich information on the preferences and the intent of the users within a short-to-medium time frame. It is a particularly important subject for the search and data mining communities because of its generality across all categories of web activities. However, there exists very little study on general non-search browsing data. Prior literature related to sessions [39, 38, 5, 6, 12] focuses almost exclusively on search trails within query sessions. However, as we will present in Section 3, search-related activities account for less than 5% of overall user activity online. Analysis of web sessions in a general setting broadens the user behavior models with richer contextual information from the entire spectrum of actions, and is key to new web search applications that aim to provide enriched user search experience centered around users' interests.

A new page importance ranking algorithm called BrowseRank [26] has recently been proposed. BrowseRank makes two significant contributions. First, it uses the more reliable input of user behavior data, computing a user browsing graph, rather than a web link graph. Second, BrowseRank models the random walk on the user browsing graph by a continuous-time Markov process. BrowseRank has shown better ranking performance compared with link analysis algorithms at the expense of higher computational costs.

Study on search trails in query sessions is the subject in a few recent works. Dramatic differences in user interaction behaviors with a search engine are reported in [39]. The idea of identifying popular end points of search trails as query-dependent feature is discussed in [38] for improving web search interaction. A recent study [5] shows improved retrieval quality using post-search browsing activities over alternative data sources that contain only the end points of search trails or clickthrough logs. Also, the study suggests that post-search browsing behavior logs provide strong signal for inferring document relevance for future queries.

3. MINING WEB SESSIONS

We define a web session as a logical unit of time-ordered user browsing activities, representing a single span of user interactions with a web browser. The concept of session in our study is generalized to all categories of web activities, while studies related to search log or search clickthrough data consider a session simply as a set of search queries and largely ignore all other activities.

3.1 Session Identification

A user's browsing history is commonly accessible from several sources, for instance, the ISP or other gateway to the web [28] or clients installed on the user's environment [39]. In this study, we use information logged by the Yahoo! Toolbar, a browser add-on that assists users with quick access to various web tasks. The toolbar logs user activities for a subset of users who opted in for this data collection during installation. Each log entry is a tuple of {*cookie*, *timestamp*, *URL*, *referral URL*, *event attribute list*}. The cookie is a unique, anonymous client identifier that expires and refreshes after a pre-defined time period. The URL is the identifier of the page being accessed and the referral URL is the URL from which the user access the current URL. The event attribute list is composed of various metadata associated with the activity. For the experiments in this paper, the browsing data consisted of over 30 billion anonymous events, across millions of unique Yahoo! users, collected over 6 months in 2008.

To segment web activities into sessions, we first use the referral URL → current URL structure to reconstruct the entire chain of browsing activities per user. This scheme ensures that, for users who are multitasking (e.g., those having multiple browser windows or tabs open), we group activities associated with different tasks into separate sessions rather than interleaving them together. Next, we partition the time-ordered user events using two boundary conditions. First, we start a new session from the current event if there

[1] While the first Google index in 1998 had 26 million pages, this number officially reached 1 trillion mark as of July 25, 2008 [15].

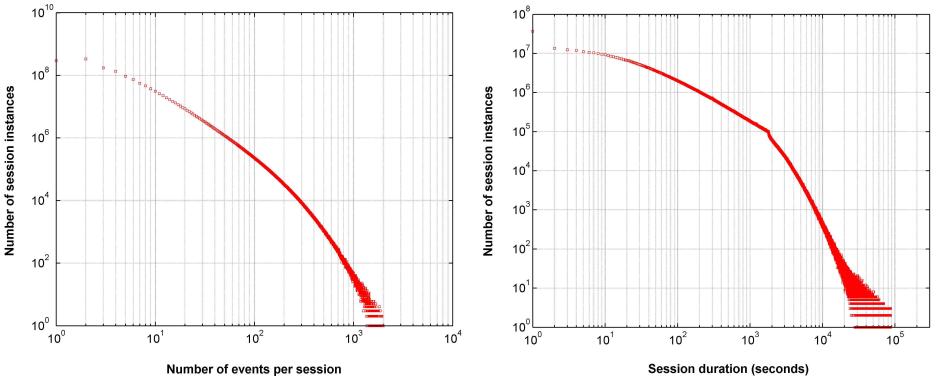


Figure 1: Distribution of session lengths (left) and session durations (right) in web-scale user browsing logs.

Table 1: Key characteristics of general web sessions.

Average events per session	9.1
Standard deviation of events per session	24.5
Average session duration (seconds)	420.3
Standard deviation of session duration (seconds)	1068.0
Sessions per user per day	15.5
Percentage of search sessions	4.85%

is more than 30 minutes of inactivity between the current event and its immediately preceding event. Second, a new session starts if the current event entry does not have a referral URL. This typically happens when the user launches a new web browser, or clicks on a link in a non-browser source (*e.g.* in a text file).

Our session segmentation approach requires only one-pass scanning over the data. While this is a simple mechanism, a recent study on finding logical sessions from query logs [6] has shown that in vast majority (92%) of the cases, a session segmentation method based on timeout threshold gives identical scores to an advanced and computationally expensive algorithm [6], when both are compared with human judged sessions using the objective Rand index [34]. For the small fraction of remaining sessions that are difficult for the advanced algorithm, the timeout-based method gives merely marginally degraded performance of 1.4%.

3.2 Session Characteristics

Table 1 summarizes the key characteristics of general web session. Figure 1 shows the probability distributions of the number of events in a session and session duration, respectively. The number of events in a web session approximately follows a power law distribution. Its mean and standard derivation are 9.1 and 24.5, respectively, demonstrating that a web session contains significantly richer activity context and diversity than a search session, which reportedly consists of an average of 5 events [5]. In addition, search sessions (those containing at least one query sent to one of the major web search engines) constituted 4.85% of overall sessions, signaling again that focusing on them may lead to a biased view in downstream analysis [30].

The session duration graph shows two different power law behaviors across the timeout threshold of 1, 800 seconds. On average, a web session lasts 420.3 seconds, with the standard deviation of 1068.0 seconds, demonstrating its short-to-medium time range coverage of user activities.

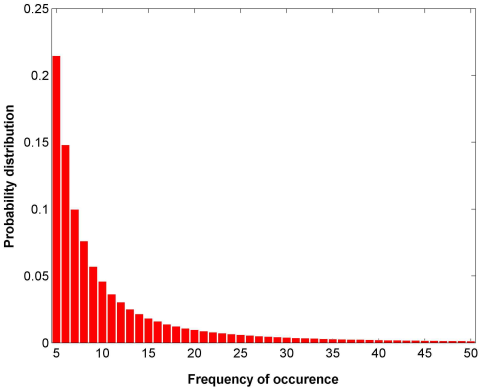


Figure 2: Probability distribution of web page occurrence in analyzed user browsing logs.

It is important to study the sparseness of content among the 30 billion events used in our study. We discover a total of 3.1 billion unique URLs. To remove individual bias, we consider web pages that are clicked by more than 5 users, which include a total of 48.5 million web pages. Figure 2 shows the distribution of web page occurrence in analyzed user browsing logs.

3.3 Session Clustering

Mining user sessions at the web scale is particularly important for learning and recognizing user behavior patterns associated with structured intents. We employ several clustering approaches to discover web sessions driven by different intents and learn their statistical characteristics. Due to space constraint, we focus our discussion on one representative clustering effort in this section.

In this experiment, we mapped each URL to an event category based on five high-level intents—search, email, information/reference, rich content (*e.g.* social networking and multimedia), and shopping. We computed the histogram representation of a session by the distribution of number of events over the intent categories. While certain temporal information is discarded, we will see in next section that this histogram representation preserves adequate discriminating power for the clustering purpose, and still being compact for the large amount of data.

Table 2: Unsupervised clustering of session histograms reveals various Web user browsing patterns. Significant features associated with each cluster are highlighted in bold.

Feature Dimension	Entire Data Mean / Standard Deviation	Cluster Centroid 1 29.8%	 2 16.6%	 3 14.3%	 4 11.9%	 5 11.0%	 6 4.7%	 7 4.6%	 8 3.5%	 9 2.1%	 10 1.5%
Search	23.63 / 37.71	0.35	**98.43**	1.19	2.35	2.33	**56.18**	**41.52**	**52.23**	6.46	0.09
Mail	16.81 / 34.98	0.07	0.66	**97.25**	0.39	0.42	1.29	**51.79**	0.70	9.79	0.07
Information	12.26 / 30.85	0.04	0.27	0.39	1.03	**96.50**	**24.58**	2.65	0.50	5.97	0.02
Content	34.31 / 45.69	**99.42**	0.37	0.64	0.45	0.36	0.64	0.95	**45.25**	**60.51**	**99.54**
Shopping	12.85 / 31.60	0.08	0.24	0.41	**95.67**	0.29	**16.92**	2.60	0.86	16.84	0.06
Events	9.06 / 24.53	11.14	2.89	5.66	6.25	5.33	4.24	5.38	4.26	7.84	**151.68**
Duration	420.30 / 1067.99	532.49	261.4	303.85	235.78	298.91	228.40	455.58	218.01	439.78	**4237.65**

To reliably associate a visit to each URL with an interpretable type, we used human categorizations of the top 1,200 most popular web sites to map events to intents. While coverage achieved this way was reasonable at 41% for all events, we augmented these categorizations using heuristics that map from URLs to likely intents; for example, URLs of the format `shopping.*.com/*` were mapped to shopping intent, and so on.

Within each session, a browsing event was labeled either as unknown, or assigned to one of these five intent categories described above. We then computed the distribution of events over the six intent labels (*i.e.*, including the unknown class), and discarded those sessions that contain more than 80% of unknown events, as they cannot be reliably clustered. Finally, we smoothed each normalized intent histogram of the remaining sessions by evenly distributing the weight associated with the unknown class to all the other five bins in the histogram.

The final session histogram is a 7 dimensional feature vector. The first 5 dimensions correspond to the normalized intent histogram, with their sum equal to 100. The last 2 dimensions correspond to the number of events in the session and the session duration in seconds, respectively.

To gain further insights on the spread in session histograms, we used principle component analysis (PCA) to reduce the dimensionality. PCA seeks projections onto a low-dimensional linear subspace that best preserves the data scatter in a least-squares sense [13]. The 3D view of session histogram shown in Figure 3 demonstrates the heterogeneity as the histogram data covers a broad continuum of activity space. Among the first 6 eigenvalues that are all significant, the first eigenvalue is dominant.

3.4 Session Interpretation

A meaningful interpretation to sessions is key to understanding the context of activities on general, unconstrained user behavior data. Table 2 summarizes the unsupervised session histogram clustering results using k-means algorithm with $k = 10$. These clusters are ordered based on the cluster size. Significant features that give clear indication of cluster attributes in Table 2 are highlighted in bold.

Various intent-driven web browsing patterns clearly emerge from statistical properties of the clusters. The top 5 clusters correspond to coherent sessions of rich content browsing, search, email, shopping, and information, respectively. For instance, the center of cluster 1, with 29.8% of entire data, contains 99.42% rich content browsing. Typically, these are interactions of users with social networking sites, such as Facebook and MySpace. Its cluster-wise standard deviation of 2.82% along this feature dimension is significantly smaller than the standard deviation of 45.69% for the entire data.

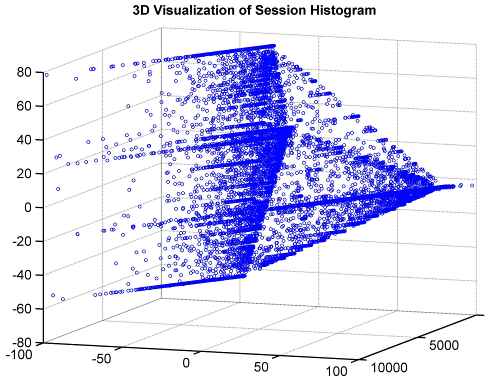


Figure 3: Visualization of session histograms in 3D by dimensionality reduction using PCA.

Clusters revealing more sophisticated user behaviors are also evident in Table 2. These interesting patterns include browsing web search results without a click (cluster 2), collecting information during shopping (cluster 6), visits to rich content web site through navigational queries (cluster 8), and prolonged activities in social networking sites (cluster 10, note the average session duration).

These observations demonstrate that even a simple approach to session representation—as distributions over high-level classes—can provide the search engine with valuable information, such as the distribution over the types of content that users are likely to access (useful for crawling scheduling as well as for ranking purposes). If we apply filter to the entire set of sessions and preserve only those containing search queries, we can further observe what queries often lead to a particular session type (e.g., a shopping session) and optimize the user experience for that.

4. USING BROWSING INFORMATION FOR WEB SEARCH

In this section, we present a novel web search ranking algorithm, ClickRank, that combines different notions of user preferences mined from browsing sessions. The ClickRank algorithm provides a robust estimate of the importance of web pages and websites without explicitly constructing a web graph; its relatively low computational cost make it particularly useful for web search ranking purposes. We also describe how ClickRank can be incorporated with a large set of other ranking features for learning a ranking model.

4.1 ClickRank

A web session contains several contextual indicators of user preferences among the visited web pages. Intuitively, users tend to browse content that they perceive as important in the context of their informational need. This makes the dwell time on a web page an important endorsement of the user's interest level in it. The click order within a general trail of user activities is also important: accessing one web page before another in the session may be interpreted as a preference signal coming from the user. ClickRank aims to combine these signals to determine a local importance value for each page within a session, and then aggregate the importance values over all sessions of interest.

We start by computing local importance values within each session. The ClickRank of a web page p_i in a given web session s_j is a function of several indicators within the session context—the dwell time on the page, the page load time, the rank of p_i in the ordered set of all visited URLs, and the frequency of occurrence in the session. We define the local ClickRank function as

$$ClickRank(p_i, s_j) = \sum_{p_i \in s_j} w_r(i, s_j) w_t(p, s_j) I(p = p_i), \quad (1)$$

where $w_r(i, s_j)$ is a weight function induced on the rank of the event i in session s_j, and $w_t(p, s_j)$ is a weight function computed from the set of temporal attributes associated with the browsing of page p_i. $I()$ denotes the indicator function.

We define the weight function $w_r()$ for an event i in rank $r(i)$ of a session s_j with a total of n_j events as

$$w_r(i, s_j) = \frac{2(n_j + 1 - r(i))}{n_j(n_j + 1)}, \quad (2)$$

where $r(i) \in \{1, \ldots, n_j\}$ and $w_r(i, s_j)$ is a monotonically decreasing function w.r.t. the rank of the event within a session i. The function choice for $w_r()$ is motivated by measurements of implicit user preference judgements through eye tracking experiments [21], which show decreasing relative attention devoted to ordered clicks in navigational and informational tasks. Note that the sum of $w_r(i, s_j)$ over a given session s_j is always equal to 1, *i.e.*, $\sum_{i=1}^{n_j} w_r(i, s_j) = 1$.

For a set of web sessions $\mathcal{S} = (s_1, \ldots, s_k)$ across users and over a period of time, we aggregate the ClickRank values as

$$ClickRank(p, \mathcal{S}) = AGGR_{s \in \mathcal{S}} \; [ClickRank(p, s)], \quad (3)$$

where $ClickRank(p, s)$ is the local ClickRank function defined in (1) given an instance of observed sessions, and $AGGR$ denotes an aggregation function, such as summation or averaging, over all sessions of interest. In the following experiments, we use summation as the aggregation function.

Finally, the ClickRank of a website w for a set of sessions $\mathcal{S}$ is simply the sum of the ClickRank values of all pages in $\mathcal{S}$ that are part of the site: $ClickRank(w, \mathcal{S}) = \sum_{p \in w} ClickRank(p, \mathcal{S})$. Note that using a sum implicitly models both the importance (as evidenced by ClickRank values of individual pages) and the size of the website – the amount of pages that it consists of.

4.2 Theoretical Analysis

Our formulation of ClickRank has a theoretical interpretation based on an intentional surfer model. A web session can be viewed as a logical sequence of hops through the hyperlink structure of the web. At each step, a user selects what she judges as most relevant as the next click, based on a variety of features such as the attractiveness of content in the context of the use's activity, her prior knowledge, and so on. The user further indicates her interest through various temporal attributes, such as the time devoted to the page or whether it was visited multiple times. This process continues throughout the duration of the session, until the user starts another journey on the web.

More concretely, the local ClickRank function defines a random variable $X_j^i : \Omega \to \mathbb{R}_0^+$ associated with the web page p_j, given an observed session s_i. X_j^i is bounded for all practical purposes, so $E(X_j^i) < \infty$ and $var(X_j^i) < \infty$. Denote the set of random variables associated with the web page p_j over the entire set of observed sessions $\mathcal{S} = (s_1, \ldots, s_k)$ by $\{X_j^1, X_j^2, \ldots, X_j^k\}$, and assume they are independent and identically distributed. As $k \to \infty$, $\frac{1}{k} \sum_{i=1}^{k} X_j^i$ converges to $E(X_j^i)$ a.s. by the strong law of large numbers.

We can establish bounds on a ClickRank-induced function in a probabilistic setting by the following theorems.

Theorem 1. *Let $f : \mathbb{R} \to [0, +\infty)$ be a non-negative function, then*

$$\mathbb{P}[f(X) \geq a] \leq \frac{E(f(X))}{a} \text{ for all } a > 0.$$

Theorem 2. *If $f : \mathbb{R} \to [0, +\infty)$ is a non-negative function taking values bounded by some number M, then*

$$\mathbb{P}[f(X) \geq a] \geq \frac{E(f(X)) - a}{M - a} \text{ whenever } 0 \leq a < M.$$

Simply put, as the volume of these web browsing sessions analyzed by ClickRank reaches a sizable sample of the entire web traffic, the rank computed by ClickRank for each page converges to its true rank according to a usage criterion.

4.3 Application to Ranking

As a query-independent feature, ClickRank can be incorporated into a document ranking process in several ways [10]. One particular framework that has recently become prominent, is the *learning to rank* approach to information retrieval, which aims to apply machine-learning algorithms to derive a ranking function from data. In a machine-learned ranking framework, a large variety of features are used to model a query and a document. Query features can be its length or frequency in a search log, and document features can be term statistics or, in the case of web documents, the number of incoming HTML links. Machine learned ranking provides a convenient approach for quantitatively evaluating the effectiveness of ClickRank as a novel feature on top of a large collection of existing ranking features.

We learn the ranking model using the functional regression framework of gradient boosting [14], which expresses the solution to the ranking function as additive expansion of M

parameterized functions

$$f^*(\mathbf{x}) = \sum_{i=0}^{M} f_m(\mathbf{x}) \equiv \sum_{i=0}^{M} \beta_m h(\mathbf{x};\mathbf{a}_m), \qquad (4)$$

where $f_0(\mathbf{x})$ is an initial guess, and $[f_m(\mathbf{x})]_1^M$ are incremental functions (or "boosts"). In Equation(4), each incremental function $f_m(\mathbf{x})$ can be further factored as the product of a base learner $h(\mathbf{x};\mathbf{a}_m)$ and corresponding coefficient β_m.

The idea of gradient boosting is to sequentially fit a parameterized function to current residuals by least-squares criterion at each iteration

$$y_{im} = -[\frac{\partial\Psi(y_i, f(\mathbf{x}_i))}{\partial\Psi(f(\mathbf{x}_i))}]_{f(\mathbf{x})=f_{m-1}(\mathbf{x})} \qquad (5)$$

and

$$\mathbf{a}_m = \arg\min_{\mathbf{a},\beta} \sum_{i=1}^{N} [y_{im} - \beta h(\mathbf{x}_i;\mathbf{a})^2], \qquad (6)$$

where N is the number of training samples. The optimal coefficient β_m is computed by line search

$$\beta_m = \arg\min_{\beta} \sum_{i=1}^{N} \Psi(y_i, f_{m-1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i;\mathbf{a}_m)). \qquad (7)$$

We use decision tree as the base learner $h(\mathbf{x};\mathbf{a}_m)$ in (4), where it is parameterized by the splitting variables and corresponding split points. At each iteration m, a decision tree partitions the entire feature space into disjoint regions $[R_{lm}]_{l=1}^L$ and predicts based on the region that contains the observed feature vector $\mathbf{x}$ as

$$h(\mathbf{x};[R_{lm}]_1^L) = \sum_{i=1}^{L} y_{lm} I(\mathbf{x} \in R_{lm}). \qquad (8)$$

Gradient boosted decision trees (GDBT) produce competitive, highly robust, interpretable procedures in regression and classification [14], and are particularly useful for settings with large amounts of data and a dense feature space.

4.4 Relation to graph-based models

ClickRank has a number of advantages over approaches that estimate the web page authority from explicit graph formulations, such as PageRank and BrowseRank. First, ClickRank is data driven, and does not embed assumptions on the traversing scheme over the web. Second, it is significantly more computationally efficient: local ClickRank values are inexpensive to calculate and can be derived independently for each session. This makes ClickRank well-suited to distributed computation (e.g., the MapReduce computation paradigm [11] that was used for the experiments in this paper), as well as memory friendly. Furthermore, addition of new data requires only incremental computation of new local ClickRank values on the newly logged web sessions, and combining with those from existing sessions, rather than recomputation of the entire model (such as would be needed by PageRank and BrowseRank). This is particularly important for the processing of web-scale user browsing information, which is constantly changing.

5. EXPERIMENTS

We demonstrate the effectiveness of ClickRank algorithm in three core aspects of web search—site ranking, page ranking, and mining new, popular web pages. In our experiments, we assume that the dwell time on a page and the page load time are two independent random processes and define the temporal weight function in (1) as

$$w_t(p,s) = (1 - e^{-\lambda_1 t_d}) e^{-\lambda_2 t_l} I(t(p) \in \mathcal{T}), \qquad (9)$$

where t_d and t_l are the normalized dwell time on the page and page load time w.r.t. the entire session. $t(p)$ is the timestamp of the event, and $\mathcal{T}$ denotes the time range.

In the following experiments, we used the same 6 months of aggregate user browsing logs collected from the Yahoo! toolbar. In total, the data contains more than 3.3 billion web sessions. These sessions contain 16.3 million unique websites, and 3.1 billion unique web pages.

Table 3: Top-ranked sites with different algorithms

Rank	PageRank	BrowseRank	ClickRank
1	adobe.com	myspace.com	yahoo.com
2	wordpress.com	msn.com	google.com
3	w3.org	yahoo.com	myspace.com
4	miibeian.gov.cn	youtube.com	live.com
5	statcounter.com	live.com	youtube.com
6	phpbb.com	facebook.com	facebook.com
7	baidu.com	google.com	msn.com
8	php.net	ebay.com	friendster.com
9	microsoft.com	hi5.com	pogo.com
10	mysql.com	bebo.com	aol.com
11	mapquest.com	orkut.com	microsoft.com
12	cnn.com	aol.com	wikipedia.org
13	google.com	friendster.com	ebay.com
14	blogger.com	craigslist.org	craigslist.org
15	paypal.com	google.co.th	hi5.com
16	macromedia.com	microsoft.com	go.com
17	jalbum.net	comcast.net	ask.com
18	nytimes.com	wikipedia.org	google.co.th
19	simplemachines.org	pogo.com	comcast.net
20	yahoo.com	photobucket.com	orkut.com

5.1 Site Ranking

We computed the ClickRank for each website and ordered them by this value. We list the top-ranking 20 sites computed with ClickRank, and compare them to those computed by PageRank[2] and BrowseRank[3]. Results are listed in Table 3, following the same convention used in [26].

On the task of site ranking, our results confirm the same finding reported in [26] that link analysis algorithms like PageRank have a strong bias towards sites with higher degree of inlinks and do not necessarily reflect the degree of actual usage. This is a fundamental limitation of the web link graph, from which PageRank and other link-based authority estimation algorithms are derived.

The computed site ranked lists by both ClickRank and BrowseRank algorithms are surprisingly similar, with a total of 18 overlapping entries among the top 20 sites. Both ranked lists correlates better with web users' informational need compared to PageRank, as they are both computed over user behavior data. Some ranking differences between BrowseRank and ClickRank in this table can be attributed to their data source. BrowseRank is computed over a set

[2] Using the web link graph as constructed by the Yahoo! crawler.

[3] This list is included from the reported list in [26] on a total of 5.6 million websites.

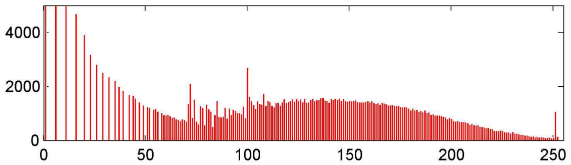


Figure 4: Distribution of discretized ClickRank score over a large collection of judged documents.

of users who installed the Live toolbar, and are presumably users of live.com and msn.com services; similarly, ClickRank is computed over a set of Yahoo! users.

One key difference between the results produced by ClickRank and BrowseRank is that ClickRank consistently ranks the starting point of user's web experience higher. One of the major search engines, `ask.com`, does not even appear among top 20 sites produced by BrowseRank.

ClickRank has a significantly lower computational cost than PageRank or BrowseRank. ClickRank requires only one pass through the data, and does not require building intermediate graphs and solving stationary probability distributions. This also allows for rapid adaptation of ClickRank values to new content: as noted earlier, new browsing information that is collected does not require recomputation over entire data. The overall running time of our implementation of ClickRank algorithm in ranking of the 16.3 million websites in this section and 3.1 billion web pages for the page ranking test in the next section are 56 minutes and 1 hour 32 minutes, respectively, using the map-reduce framework on 300 Hadoop nodes. To our best knowledge, these are the best published run times for page importance ranking on a web scale.

In a realistic, production-grade search engine environment, it is important to minimize the footprint of every relevance feature used by the ranking model so that latency and memory requirements are met. Often, float numeric values are compressed or discretized into a small dynamic range that can be represented with as few bits as possible. To this end, and to evaluate the ranking performance of ClickRank as deployed in a production system, we quantize the computed ClickRank score for each web page into an unsigned byte within the range of $[0, \dots, 255]$. The distribution of these values are shown in Figure 4.

5.2 Page Ranking

5.2.1 Evaluation Methodology

We comprehensively evaluated the page ranking performance of ClickRank in conjunction with several hundred additional features used in commercial search engines. To gain further insights, we quantified the search improvement from ClickRank with a state-of-the-art baseline system, and measured its relative variable significance against this large pool of ranking features. This evaluation scheme gives more realistic, quantitative results in contrast to common published evaluations using limited feature set as baselines. For instance, [26] employs the single feature of BM25 [22] as the relevance baseline in their evaluation.

We used discounted cumulative gain (DCG) and normalized discounted cumulative gain (NDCG), two widely used search engine relevance measures [18, 19], to quantitatively evaluate ranking performances. Given a query and a ranked list of returned documents in response to the query, the $DCG(K)$ score for the query is computed as

$$DCG(K) = \sum_{k=1}^{K} \frac{g_k}{\log_2(1+k)}, \qquad (10)$$

where g_k is the weight for the document at rank k. A five-grade score is assigned to each document based on its degree of relevance.

We trained ranking models using gradient boosted decision trees on the baseline system with all existing features, and on the alternative system that includes one additional ClickRank feature. Training and test data is partitioned through cross-validation. We used identical parameter settings in all the following comparison experiments.

To quantify the relative importance S_i of each individual input ranking feature x_i, we used the following measure of variable importance for decision trees [7]

$$S_i^2 = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{L-1} \frac{w_l w_r}{w_l + w_r} (\overline{y}_l - \overline{y}_r)^2 I(v_t = i), \qquad (11)$$

where v_t is the splitting variable at the non-terminal node n, $\overline{y}_l$, $\overline{y}_r$ are the means of the regression responses in the left and right subtrees respectively, and w_l, w_r are the corresponding sums of the weights.

5.2.2 Data Preparation

In this experiment, we used a set of randomly sampled $9,041$ queries from a search log. For each query, 5–20 web pages have been independently judged by a panel of editors and assigned with one of the five relevance scores.

5.2.3 Results and Discussion

The usage of ClickRank as an additional relevance feature brings 1.02%, 0.97%, 1.11%, and 1.331% web search improvements in $DCG(1)$, $DCG(5)$, $DCG(10)$, and $NDCG$, respectively, on top of a state-of-the-art ranking model—a model already incorporating hundreds of features derived from content (*e.g.*, anchor, title, body, and section), from the link structure of the web, from search engine query logs, and from other sources. The reported gains are strongly statistically significant.

The gains in retrieval performance from ClickRank on top of a competitive search engine are summarized in Table 4. These improvements are very substantial in the context of commercial web search: our strong baseline incorporates a feature set of several hundred signals tuned over a long period of time. In addition, 81.2% out of over $9,000$ queries are affected in the alternative experiment, demonstrating the generality of ClickRank. Furthermore, we observe higher improvements on long queries in Table 4, which are typically much more challenging for search engines. We show search improvements across different query lengths in Figure 5.

We experimented with several variants of ClickRank, and observed that it consistently ranks among the top features in variable significance as calculated by (11). For example, in the experiment shown in Figure 5 and Table 4, the ClickRank feature is ranked 25th in variable importance among several hundred other features, significantly higher than the highest-ranking feature derived from page visit counts (ranked 56th) and a feature based on a propagation of authority through the link graph (ranked 108th).

Table 4: ClickRank delivers statistically significant web search improvements over a state-of-the-art baseline.

Query Length	Number of Queries	Affected Queries	Improvements in $DCG(1)$	$DCG(5)$	$DCG(10)$	$NDCG$	Significance Test p-value
1	1484	1232	0.448%	0.713%	1.002%	0.378%	5.33×10^{-2}
2	2992	2450	0.560%	0.993%	1.121%	1.071%	4.65×10^{-4}
3	2153	1722	1.618%	1.076%	1.406%	2.177%	1.10×10^{-4}
4+	2412	1937	0.918%	0.861%	0.784%	1.433%	1.61×10^{-5}
All	9041	7341	1.020%	0.966%	1.105%	1.331%	9.98×10^{-5}

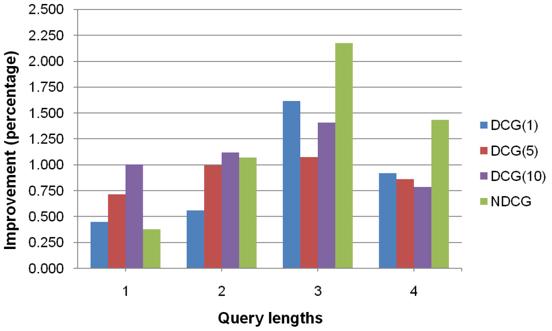


Figure 5: Search improvements from ClickRank for different query lengths.

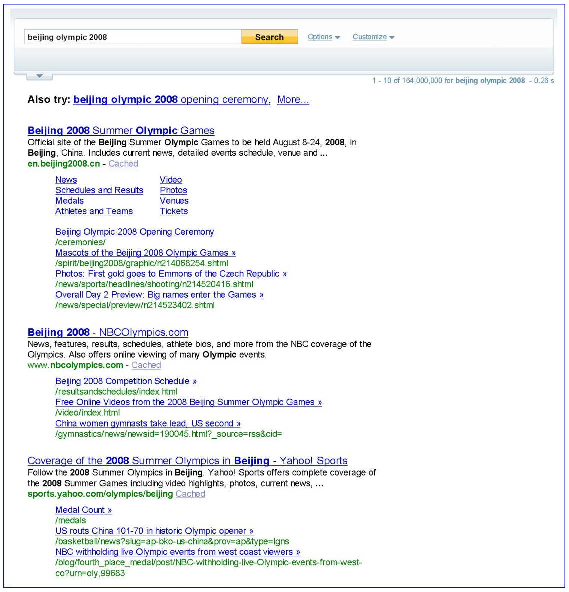
These results demonstrate the significance of session-based web importance estimation and show that ClickRank captures novel user preference knowledge not identified through other modeling techniques.

5.3 Mining Dynamic Quicklinks

Many web search engines supply a set of "quicklinks" – direct access links to certain pages within the site, in addition to the search result itself. For example, while the top result for the navigational query "ebay" is www.ebay.com, it also contains quick access links to popular destination within ebay.com, such as "Motors" (links to motors.ebay.com), "Half Books" (links to half.ebay.com), and so on. Typically, these quicklinks are pointers to frequently visited destinations within the host mined from query or clickthrough logs. This method, however, has two major limitations. First, query logs do not contain user activities beyond the scope of interactions with search engines, which account for the vast majority (more than 95%, as we showed earlier) of real web traffic. Second, results computed from query logs have a strong bias towards old, navigational links within the site since they receive more clicks within the visibility range of search engines.

We demonstrate a novel application of ClickRank for discovering and displaying dynamic quicklinks in web search results through recency ranking. The idea is to adapt the time range for the indictor function in Equation (9) w.r.t. the content refresh rate found by web crawlers. In addition to normal search results, the system displays highly ranked web pages computed by ClickRank as quicklinks to the user.

Figure 6 shows search results with discovered quicklinks by the system in response to the query of "beijing olympic 2008" on two days during summer Olympic Games 2008, using the time range of 24 hours. Quicklinks mined by ClickRank are displayed side by side with the most frequently



1 - 10 of 164,000,000 for **beijing olympic 2008** - 0.26 s

Also try: **beijing olympic 2008** opening ceremony, More...

Beijing 2008 Summer **Olympic** Games
Official site of the **Beijing** Summer **Olympic** Games to be held August 8-24, **2008**, in **Beijing**, China. Includes current news, detailed events schedule, venue and ...
en.beijing2008.cn - Cached

News	Video
Schedules and Results	Photos
Medals	Venues
Athletes and Teams	Tickets

Beijing Olympic 2008 Opening Ceremony
/ceremonies/
Mascots of the Beijing 2008 Olympic Games »
/spirit/beijing2008/graphic/n214068254.shtml
Photos: First gold goes to Emmons of the Czech Republic »
/news/sports/headlines/shooting/n214520416.shtml
Overall Day 2 Preview: Big names enter the Games »
/news/special/preview/n214523402.shtml

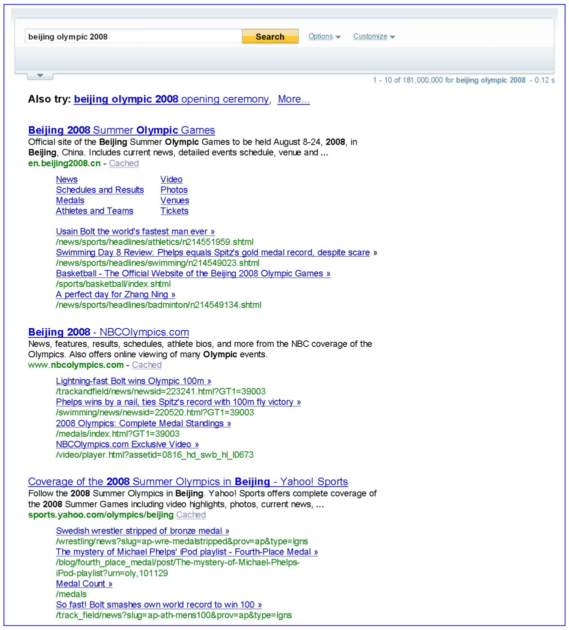
Beijing 2008 - NBCOlympics.com
News, features, results, schedules, athlete bios, and more from the NBC coverage of the Olympics. Also offers online viewing of many **Olympic** events.
www.**nbcolympics.com** - Cached

Beijing 2008 Competition Schedule »
/resultsandschedules/index.html
Free Online Videos from the 2008 Beijing Summer Olympic Games »
/video/index.html
China women gymnasts take lead, US second »
/gymnastics/news/newsid=190045.html?_source=rss&cid=

Coverage of the **2008** Summer Olympics in **Beijing** - Yahoo! Sports
Follow the **2008** Summer Olympics in **Beijing**. Yahoo! Sports offers complete coverage of the **2008** Summer Games including video highlights, photos, current news, ...
sports.yahoo.com/olympics/beijing Cached

Medal Count »
/medals
US routs China 101-70 in historic Olympic opener »
/basketball/news?slug=ap-bko-us-china&prov=ap&type=lgns
NBC withholding live Olympic events from west coast viewers »
/blog/fourth_place_medal/post/NBC-withholding-live-Olympic-events-from-west-co?urn=oly,99683

(a) Search results with quicklinks for August 10, 2008



1 - 10 of 181,000,000 for **beijing olympic 2008** - 0.12 s

Also try: **beijing olympic 2008** opening ceremony, More...

Beijing 2008 Summer **Olympic** Games
Official site of the **Beijing** Summer **Olympic** Games to be held August 8-24, **2008**, in **Beijing**, China. Includes current news, detailed events schedule, venue and ...
en.beijing2008.cn - Cached

News	Video
Schedules and Results	Photos
Medals	Venues
Athletes and Teams	Tickets

Usain Bolt the world's fastest man ever »
/news/sports/headlines/athletics/n214551959.shtml
Swimming Day 8 Review: Phelps equals Spitz's gold medal record, despite scare »
/news/sports/headlines/swimming/n214549023.shtml
Basketball - The Official Website of the Beijing 2008 Olympic Games »
/sports/basketball/index.shtml
A perfect day for Zhang Ning »
/news/sports/headlines/badminton/n214549134.shtml

Beijing 2008 - NBCOlympics.com
News, features, results, schedules, athlete bios, and more from the NBC coverage of the Olympics. Also offers online viewing of many **Olympic** events.
www.**nbcolympics.com** - Cached

Lightning-fast Bolt wins Olympic 100m »
/trackandfield/news/newsid=223241.html?GT1=39003
Phelps wins by a nail, ties Spitz's record with 100m fly victory »
/swimming/news/newsid=220520.html?GT1=39003
2008 Olympics: Complete Medal Standings »
/medals/index.html?GT1=39003
NBCOlympics.com Exclusive Video »
/video/player.html?assetid=0816_hd_swb_hl_l0673

Coverage of the **2008** Summer Olympics in **Beijing** - Yahoo! Sports
Follow the **2008** Summer Olympics in **Beijing**. Yahoo! Sports offers complete coverage of the **2008** Summer Games including video highlights, photos, current news, ...
sports.yahoo.com/olympics/beijing Cached

Swedish wrestler stripped of bronze medal »
/wrestling/news?slug=ap-wre-medalstripped&prov=ap&type=lgns
The mystery of Michael Phelps' iPod playlist - Fourth-Place Medal »
/blog/fourth_place_medal/post/The-mystery-of-Michael-Phelps-iPod-playlist?urn=oly,101129
Medal Count »
/medals
So fast! Bolt smashes own world record to win 100 »
/track_field/news?slug=ap-ath-mens100&prov=ap&type=lgns

(b) Search results with quicklinks for August 16, 2008

Figure 6: Dynamic quicklinks discovered using ClickRank by recency ranking.

clicked navigational links. The quicklinks effectively capture the event highlights, while the most frequently clicked navigational links remain unchanged. The quicklink results by ClickRank are more meaningful in suggesting content that are of potential interest to web users, than those that reflect the structural property of the website.

6. CONCLUSIONS

In this paper, we explored the direction of mining general user browsing information for discovering session models driven by structured user intents, and proposed user preference models that incorporate rich session context for web search ranking. We presented characteristics of general web browsing sessions and revealed interesting user behavior patterns mined from sessions. We introduced *ClickRank*, an efficient, scalable algorithm for estimating web page and website importance based on user preference judgments mined from session context. ClickRank is based on a data-driven intentional surfer model, and is empirically shown to be an effective and novel ranking feature even on top of a highly competitive baseline system employing hundreds of ranking features. We also discussed the advantages of ClickRank over existing importance ranking approaches. ClickRank is effective and efficient to compute, delivering highly correlated ranking results compared to the state-of-the-art approach that utilize browsing data. We also demonstrated a promising application that mines dynamic quicklinks for enhancing web user experience.

These promising results, together with earlier findings from user browsing data in web search trails, highlight the great potential of data-driven user behavior modeling at the web scale. As the amount of web traffic continues to grow exponentially, we expect that explicit information about user behavior online will play an increasingly prominent role in web search and in the modeling of user intents.

7. ACKNOWLEDGMENTS

We thank Benoit Dumoulin and other members of the Query Intent Modeling Team in Yahoo! Search and Yahoo! Research Barcelona for useful discussions. The authors would also like to thank Emre Velipasaoglu, Scott Farrar, Jerry Ye, and the Yahoo! grid team for their support.

8. REFERENCES

[1] E. Adar, D. S. Weld, B. N. Bershad, and S. S. Gribble. Why we search: visualizing and predicting user behavior. In *WWW*, pages 161–170, 2007.

[2] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR*, pages 19–26, 2006.

[3] E. Agichtein and Z. Zheng. Identifying "best bet" web search results by mining past user behavior. In *KDD*, pages 902–908, 2006.

[4] R. Baeza-Yates, C. Castillo, F. Junqueira, V. Plachouras, and F. Silvestri. Challenges in distributed information retrieval. In *ICDE*, pages 6–20, 2007.

[5] M. Bilenko and R. W. White. Mining the search trails of surfing crowds: Identifying relevant websites from user activity. In *WWW*, pages 51–60, 2008.

[6] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: Model and applications. In *CIKM*, pages 609–618, 2008.

[7] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. CRC Press, 1984.

[8] A. Z. Broder, R. Lempel, F. Maghoul, and J. Pedersen. Efficient PageRank approximation via graph aggregation. In *WWW*, pages 484–485, 2004.

[9] W. Cohen, R. Shapire, and Y. Singer. Learning to order things. *Journal of Artificial Intelligence Research*, 10:243–270, 1999.

[10] N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *SIGIR*, pages 416–423, 2005.

[11] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In *OSDI*, pages 137–150, 2004.

[12] D. Downey, D. Liebling, and S. Dumais. Understanding the relationship between searchers' queries and information goals. In *CIKM*, pages 449–458, 2008.

[13] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2 edition, 2000.

[14] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[15] Google. We know the web was big. Online, 2008. `http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html`.

[16] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *VLDB*, pages 576–587, 2004.

[17] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang. Accessing the deep Web. *Communications of the ACM*, 50(5):94–101, 2007.

[18] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *SIGIR*, pages 41–48, 2000.

[19] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[20] T. Joachims. Optimizing search engines using clickthrough data. In *KDD*, pages 133–142, 2002.

[21] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR*, pages 154–161, 2005.

[22] K. S. Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: Development and comparative experiments (parts 1 and 2). *Information Processing and Management*, 36(6):779–840, 2000.

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

[24] A. N. Langville and C. D. Meyer. Deeper inside PageRank. *Journal of Internet Mathematics*, 1(3):335–400, 2005.

[25] X. Li, Y.-Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In *SIGIR*, pages 339–346, 2008.

[26] 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 *SIGIR*, pages 451–458, 2008.

[27] F. McSherry. A uniform approach to accelerated PageRank computation. In *WWW*, pages 575–582, 2005.

[28] M. R. Meiss, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani. Ranking web sites with real user traffic. In *WSDM*, pages 65–76, 2008.

[29] C. Moler. The world's largest matrix computation. Online, 2002. `http://www.mathworks.com/company/newsletters/news_notes/clevescorner/oct02_cleve.html`.

[30] A. Mowshowitz and A. Kawaguchi. Bias on the Web. *Communications of the ACM*, 45(9):56–60, 2002.

[31] C. Olston and S. Pandey. Recrawl scheduling based on information longevity. In *WWW*, pages 437–446, 2008.

[32] L. Page, S. Brin, R. Motwani, and T. Winograd. *The PageRank Citation Ranking: Bringing Order to The web*. Technical Report, Stanford University, 1998.

[33] B. Piwowarski and H. Zaragoza. Predictive user click models based on click-through history. In *CIKM*, pages 175–182, 2007.

[34] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[35] D. E. Rose and D. Levinson. Understanding user goals in web search. In *WWW*, pages 13–19, 2004.

[36] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. *ACM SIGIR Forum*, 33(1):6–12, 1999.

[37] B. Tan, X. Shen, and C. Zhai. Mining long-term search history to improve search accuracy. In *KDD*, pages 718–723, 2006.

[38] R. W. White, M. Bilenko, and S. Cucerzan. Leveraging popular destinations to enhance web search interaction. *ACM Trans. Web*, 2(3):1–30, 2008.

[39] R. W. White and S. M. Drucker. Investigating behaviorial variability in web search. In *WWW*, pages 21–30, 2007.