Dr. Searcher and Mr. Browser: A Unified Hyperlink-Click Graph

Barbara Poblete[1]
barbara.poblete@upf.edu

Carlos Castillo[2]
chato@yahoo-inc.com

Aristides Gionis[2]
gionis@yahoo-inc.com

[1]University Pompeu Fabra
Barcelona, Spain

[2]Yahoo! Research
Barcelona, Spain

ABSTRACT

We introduce a unified graph representation of the Web, which includes both structural and usage information. We model this graph using a simple union of the Web's hyperlink and click graphs. The hyperlink graph expresses link structure among Web pages, while the click graph is a bipartite graph of queries and documents denoting users' searching behavior extracted from a search engine's query log.

Our most important motivation is to model in a unified way the two main activities of users on the Web: *searching* and *browsing*, and at the same time to analyze the effects of random walks on this new graph. The intuition behind this task is to measure how the combination of link structure and usage data provide additional information to that contained in these structures independently.

Our experimental results show that both hyperlink and click graphs have strengths and weaknesses when it comes to using their stationary distribution scores for ranking Web pages. Furthermore, our evaluation indicates that the unified graph always generates consistent and robust scores that follow closely the best result obtained from either individual graph, even when applied to "noisy" data. It is our belief that the unified Web graph has several useful properties for improving current Web document ranking, as well as for generating new rankings of its own. In particular stationary distribution scores derived from the random walks on the combined graph can be used as an indicator of whether structural or usage data are more *reliable* in different situations.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**];
H.2.8 [**Information Systems**]: Data Mining

General Terms

Algorithms, Experimentation, Human Factors

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.

Keywords

Search Engine Queries, Usage Mining, Structure Mining, Random-Walks, Web Graphs

1. INTRODUCTION

In recent years, significant amount of research has been devoted to studying the *Web graph* (which we refer to as *hyperlink graph* to avoid ambiguity) and the *click graph*. The hyperlink graph is the directed graph among Web pages in which edges represent hyperlinks. The click graph is a view of the information contained in query logs, i.e., a bipartite graph between queries and Web pages, in which edges connect a query with the documents that were clicked by users as a result.

At an intuitive level, these two graphs capture two of the most common tasks of users on the Web: *browsing* and *searching*. A user who browses the Web essentially follows edges on the hyperlink graph, while a user who searches and consequently clicks on the result pages, is following edges on the click graph. Searching and browsing together are equivalent to the two prototypical actions of information seeking and exploration.

The edges of these two graphs can capture certain semantic relations between the objects they represent. An example of such a relation is *similarity*: two pages connected together by a hyperlink, or a query and a page connected together by a click, are more likely to be similar than two non-connected objects [7]. Another presumed semantic relation is *authority endorsement*: a hyperlink from a page u to a page v, or a click from a query q to the page v, can both be viewed as implicit "votes" for page v [14]. These hypotheses provide a foundation for the research of several Web information retrieval problems, for instance clustering of Web pages, queries and users, on-line community discovery. Similarity search exploits the similarity hypotheses, while ranking leverages the authority-endorsement theories.

Unfortunately both the hyperlink graph and the click graph have certain disadvantages. For example, Google's PageRank [4] uses links in the hyperlink graph to compute importance scores for Web pages. As a result substantial adversarial effort has been put into artificially increasing the PageRank score of Web pages. This adversarial effort takes the form of *spam* pages or *link farms* [10, 9].

Similarly, the click graph has its own disadvantages. One of those disadvantages is its sparsity: a page that is clicked for a certain query must first appear in the list of results for that query. This may not be trivial considering the vast number of pages available for each query. Also, there is an

issue of an inherent bias in any rankings produced by this graph, favoring already highly ranked Web pages. Another related problem is its large dependency on textual matching: typically search engines emphasize precision at the expense of recall, and display only results which match exactly all the query terms, causing many relevant pages not to be connected with queries if they are not exact matches. Furthermore, the click graph is also prone to spam, but in this case *click spam* which aims towards taking advantage of usage mining algorithms to improve search ranking.

Contributions of this work. In this paper we propose a new type hybrid Web graph, which combines the existing hyperlink and the click graphs, and we apply Web mining and link-analysis algorithms to it. This new graph, which we call the *hyperlink-click graph*, is a simple graph union: it has two types of nodes, *pages* and *queries*, with directed edges between pages according to the hyperlink graph, and undirected edges between queries and pages according to the click graph.

The union of these two graphs combines the traditional hyperlink graph, based on connectivity structure, and the click graph, based on search engine usage information. The purpose of this graph is to extend the traditional hyperlink graph into a graph which reflects more accurately users' natural behavior in the Web.

In particular we define and study random walks on the unified graph. We show that ranking according to the scores obtained from the hyperlink-click graph is similar to ranking using the score of the non-combined graph with the highest performance. The unified graph compensates where either the hyperlink or click graph execute poorly, being overall more robust and fail-safe. It is important to note that in modern Web search engines, link analysis scores in the style of PageRank might be only small components of the overall ranking function. Nevertheless, we compare directly to those scores in order to isolate the effect of the hyperlink-click graph.

Combining usage and content information in one structure can improve the quality of many Web mining algorithms. From our point of view, the two graph structures are complementary and each of them can be used to alleviate the shortcomings of the other. For example, using clicks to include user feedback on the Web graph improves its resistance against link-spam. On the other hand, by considering hyperlinks and browsing patterns we increase the density and connectivity of the click graph, and we can account for pages that users might visit *after* issuing particular queries.

Applications of the hyperlink-click graph. There are several Web mining tasks in which the hyperlink-click graph can be used:

- **Ranking of documents.** A random walk on the hyperlink-click graph can be used to obtain importance scores for documents, which can be used to enhance document ranking. This particular application is in the focus of this paper.
- **Query ranking and query recommendation.** As a by-product of the random walk on the hyperlink-click graph, importance scores are obtained not only for documents but also for queries. Such query scores can be used for query recommendation: given a query, we can use the graph to find other similar queries, and then use the importance scores to rank those queries and provide alternative query recommendations to the user.
- **Similarity search.** There have been many notions of distances among documents and among queries, which have been based on the topology of the hyperlink graph (e.g. SimRank [11]) and the click graph (e.g. [5]). Such distance functions provide building boxes for designing meaningful similarity-search algorithms. We believe that refining such graph-based distance measures for the hyperlink-click graph can lead to better notions of similarity, since the hyperlink-click graph provides richer information about the objects that it relates. These similarity metrics can be used to find communities on the Web.
- **Spam detection.** Link-based features extracted from the hyperlink graph, can be used to improve content-based spam detection algorithms [2]. It is reasonable to hypothesize that link features extracted from the hyperlink-click graph can be useful to further improve spam detection.

We plan to investigate some of these applications in future work. The main focus of this paper is the first application: enhancing the ranking of Web documents.

Roadmap. The rest of the paper is organized as follows. In Section 2 we present the related work. In Section 3 we introduce our notation and provide a formal description of the graphs used in this paper. Section 4 discusses the random walk model, which is mainly used for ranking. In Section 5 we discuss our experimental results, and finally, in Section 6 summarizes our results and conclusions.

2. RELATED WORK

The Web is an extremely rich and highly interconnected source of information, which makes Web mining a very active research field. Given the space limitations, our coverage of the topic is by no means complete.

In general, the information found on the Web can be analyzed from three main points of view associated to the predominant types of data found in it [18].

Content: The information that the Web documents were designed to convey. This data consists mainly of text and multimedia.

Structure: The description of the organization of the content within the Web. This includes mainly the hyperlink structure connecting documents and how they are organized in logical structures such as Web sites.

Usage: This data describes the history of usage of a Web site or search engine. This includes click through information, as well as queries submitted by users to search engines. This data is stored in the Web server's access logs, as well as in logs for specific applications.

There are several models for representing the information on the Web. The most popular view is the one based on structure. This approach sees the Web as a graph in which documents are nodes that are connected to each other when there is at least one hyperlink from one document to the other. This graph structure has been exploited by link-based ranking algorithms such as [4] and [12]. Both methods

rank pages according to their *importance* and *authority*, estimated by analyzing the endorsements or links from other documents.

In the work presented in [1] there is an overview of many other possible graph-based representations based on the content and usage data found on the Web. The focus is on the analysis of queries from search engines and their semantic relations, as well as their relations given by the clicks on common documents. Relations between queries can be inferred from common keywords or common clicked documents. In a similar way, relations between documents can be found by looking at shared links or words. The incorporation of document contents into these types of graphs is introduced from the words in queries, their selected documents, and also by the relations induced among documents with similar words.

With respect to usage data, a common model for query logs from search engines is in the form of a bipartite undirected graph. This graph includes two types of nodes: queries and documents. Links between the two types of nodes are generated by user clicks from queries to documents in the process of selecting a search result. This type of representation was presented in [3] and used for agglomerative clustering to find related queries and documents. Later, this view was expanded in [5] where weights were added to the undirected edges, based on the number of clicks from the query to a document. This graph is referred to as *click graph*. They study the effect of forward and backward random walks on this model for document ranking. They discuss that queries should be considered as *soft* relevance judgments, and that query logs give noisy and sparse data. The work of [5] suggest that an effective method is a backward random walk.

On the other hand, the notion of *unification* of different Web data sources is not a new one. In [19] a framework is proposed for link analysis. This framework allows to model inter-type and intra-type links between different Web objects. They discuss that any link-based model can be studied within their framework and they focus their work on users and their browsing behavior. In particular they apply this to extend the HITS algorithm by incorporating users browsing patterns.

Noise and malicious manipulation of Web content affect both the click graph and hyperlink graphs. The most typical type of manipulation is link spam on the hyperlink graph [10, 9]. In this approach artificial links are created to induce higher link-based ranks on documents. In a similar way, click graph manipulation can be produced from artificial clicks on search engine results [16, 9]. The aim of this attack is to manipulate learned ranking functions that are based on click through information. Another type of *noise* that can be found in click through data is the bias of clicks due to the position of the search result. This bias has been studied and modeled, e.g. by [8, 6].

Another perspective on query logs is to avoid considering queries individually, but use them as sequences of actions. This is explored in [17] and serves a dual purpose: it reduces the noise due to single queries, and it allows the connection of different actions of users over time.

3. WEB GRAPHS

In this section we describe three types of Web graphs: the hyperlink graph, the click graph, and the hyperlink-click graph. We introduce the notation that is used in the paper, and describe the random walks that are performed over the graphs.

The hyperlink graph. Given a set of N Web documents D we consider the *hyperlink graph* $G_{\mathrm{H}} = (D, H)$ as a directed graph, where there is an edge $(u, v) \in H$ if and only if document u has a hyperlink to document v, for $u, v \in D$.

For a document $u \in D$, the set of *in-neighbors* of u (the documents that point to u) and the set of *out-neighbors* of u (the documents that are pointed to by u) are denoted by $N_{\mathrm{IN}}(u)$ and $N_{\mathrm{OUT}}(u)$, respectively. That is, $N_{\mathrm{IN}}(u) = \{v \in D \mid (v, u) \in H\}$ and $N_{\mathrm{OUT}}(u) = \{v \in D \mid (u, v) \in H\}$. For $u \in D$, $d_{\mathrm{IN}}(u) = |N_{\mathrm{IN}}(u)|$ is the *in-degree* of document u, and $d_{\mathrm{OUT}}(u) = |N_{\mathrm{OUT}}(u)|$ is its *out-degree*.

The click graph. Let $Q = \{q_1, \ldots, q_M\}$ be the set of M unique queries submitted to a search engine during a specific period of time. In practice, in order to construct the set of unique queries we assume some simple normalization, such as normalizing for space, letter case, and ordering of the query terms. For a query $q \in Q$ we denote by $f(q)$ the *frequency* of the query q, that is, how many times the query was submitted to the search engine.

In a large-scale search engine query log, in addition to the information about which queries have been submitted, there is information about which documents are clicked by the users who submit those queries. Let $D = \{d_1, \ldots, d_N\}$ be the set of N Web documents clicked for those queries.

The click graph $G_{\mathrm{C}} = (Q \cup D, C)$ is an undirected bipartite graph that involves the set of queries Q, the set of documents D, and a set of edges C. For $q \in Q$ and $d \in D$, the pair (q, d) is an edge of C if and only if there is a user who clicked on document d after submitting the query q. The obvious prerequisite is that the document d is in the set of results computed by the search engine for the query q. To each edge $(q, d) \in C$ we associate a numeric weight $c(q, d)$ that measures the number of times the document d was clicked when shown in response to the query q.

As before, we define $N(q) = \{a \mid (q, a) \in C\}$ the set of neighboring documents of a query $q \in Q$, and $N(a) = \{q \mid (q, a) \in C\}$ the set of neighboring queries of a document $a \in D$. We then define the weighted degree of a query $q \in Q$ as $d(q) = \sum_{a \in N(q)} c(q, a)$, and similarly, the weighted degree of a document $a \in D$ as $d(a) = \sum_{q \in N(a)} c(q, a)$.

The hyperlink-click graph. Quite simply, the hyperlink-click graph G_{HC} can be seen as the *union* of the hyperlink graph and the click graph. There is a directed edge of weight 1 between documents u and v if there is a hyperlink from u to v, and there is an undirected weighted edge between query q and document d if there are clicks from q to d, and the weight of the edge is equal to the number of clicks $c(q, d)$.

4. RANDOM WALKS ON WEB GRAPHS

Given a graph $G = (V, E)$ a *random walk* on G is a process that starts at a node $v_0 \in V$ and proceeds in discrete steps by selecting randomly a node of the neighbor set of the node at the current step. A random walk on a graph of N nodes can be fully described by an $N \times N$ matrix $\mathbf{P}$ of *transition probabilities*. The i-th row and the i-th column of $\mathbf{P}$ correspond both to the i-th node of the graph, $i = 1, \ldots, N$. The $\mathbf{P}_{ij}$ entry of $\mathbf{P}$ is the probability that the next node will be the node j given that the current node is the node i.

Thus, all rows of $\mathbf{P}$ sum to 1, and $\mathbf{P}$ is called *row-stochastic* matrix.

Under certain conditions (irreducibility, finiteness, and aperiodicity, see [15] for definitions and more details) a random walk is characterized by a steady-state behavior, which is known as the *stationary distribution* of the random walk. Formally, the stationary distribution is described by an N-dimensional vector $\boldsymbol{\pi}$ that satisfies the equation $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$. Alternatively, the i-th coordinate $\boldsymbol{\pi}_i$ of the stationary-distribution vector π measures the frequency in which the i-th node of the graph is visited during the random walk, and thus, it has been used as an intuitive measure of the *importance* of each node in the graph.

Next we will consider random walks in the three different graphs we have introduced: the hyperlink graph, the click graph, and the hyperlink-click graph. We will denote the stationary distributions in those three graphs by $\boldsymbol{\pi}_{\mathrm{H}}$, $\boldsymbol{\pi}_{\mathrm{C}}$, and $\boldsymbol{\pi}_{\mathrm{HC}}$, respectively. We will refer to the values of the stationary distribution vectors as *scores*.

Random walk on the hyperlink graph. The random walk on the hyperlink graph corresponds to surfing the Web by following hyperlinks at random from the current Web page. The concept has been popularized through the seminal paper by Brin and Page [4], and its application to the Google search engine. The stationary distribution is also known as the *PageRank* vector. In the PageRank model, a step of following a random hyperlink is performed with probability α, while the walk "jumps" ("teleports" or "resets") to a random page with probability $1-\alpha$. Additionally, special care is taken when reaching a *dangling node*, a node with no outgoing edges. A common assumption is that upon reaching to a dangling node the random walk continues by selecting a target node uniformly at random. Consequently, if $\mathbf{A}_{\mathrm{H}}$ is the adjacency matrix of the Web graph G_{H}, define $\mathbf{N}_{\mathrm{H}}$ to be the normalized version $\mathbf{A}_{\mathrm{H}}$ so that all rows sum to 1. Assume that $\mathbf{N}_{\mathrm{H}}$ is defined to take care of the dangling nodes, so that if a row of $\mathbf{A}_{\mathrm{H}}$ has all 0s, then the corresponding row of $\mathbf{N}_{\mathrm{H}}$ has all values equal to $1/N$. Finally, let $\mathbf{1}_{\mathrm{H}}$ be a matrix that has the value $1/N$ in all of its entries. Then the transition-probability matrix $\mathbf{P}_{\mathrm{H}}$ of the random walk on the Web graph is given by $\mathbf{P}_{\mathrm{H}} = \alpha\mathbf{N}_{\mathrm{H}} + (1-\alpha)\mathbf{1}_{\mathrm{H}}$.

In addition to yielding a better model of surfing the Web graph, performing the random jumps with probability $(1-\alpha) \neq 0$ ensures the sufficient conditions for the stationary distribution to be defined.

Random walk on the click graph. Random walk on the click graph is similar, except for the fact that the click graph is bipartite and undirected. Being bipartite creates periodicity in the random walk, while being undirected has the consequence that the stationary distribution is proportional to the degree of each node. However, assuming that we also perform random jumps with probability $(1-\alpha)$, then the random walk is aperiodic and irreducible (every node can be reached from every other node), and also the stationary distribution at each node is not a direct function of its degree.

Formally, the random walk on the click graph is described as follows. Let $\mathbf{A}_{\mathrm{C}}$ be an $M \times N$ matrix, whose M rows correspond to the queries of Q and the N columns correspond to the documents of D, and whose (q,d) entry has value $c(q,d)$, the number of clicks between query $q \in Q$ and document $d \in D$. Let $\mathbf{A}'_{\mathrm{C}}$ be an $(M+N)\times(M+N)$ matrix defined by

$$\mathbf{A}'_{\mathrm{C}} = \begin{pmatrix} \mathbf{A}_{\mathrm{C}} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{\mathrm{C}}^T \end{pmatrix},$$

and let $\mathbf{N}_{\mathrm{C}}$ be the row-stochastic version of $\mathbf{A}'_{\mathrm{C}}$. Here again we assume that $\mathbf{N}_{\mathrm{C}}$ is defined to take care of the dangling nodes, so that if a row of $\mathbf{A}_{\mathrm{C}}$ has all 0s, then the corresponding row of $\mathbf{N}_{\mathrm{C}}$ has all values equal to $1/(M+N)$. Finally, let $\mathbf{1}_{\mathrm{C}}$ be an $(M+N)\times(M+N)$ matrix that has value $1/(M+N)$ in all its entries. Then the transition-probability matrix that describes the random walk on the click graph is $\mathbf{P}_{\mathrm{C}} = \alpha\mathbf{N}_{\mathrm{C}} + (1-\alpha)\mathbf{1}_{\mathrm{C}}$.

Note that in [5] a backward random walk is used, while we consider instead a forward random walk.

Random walk on the hyperlink-click graph. Using the notation that we introduced in the previous paragraphs, the random walk on the hyperlink-click graph is defined as follows: First overwrite $\mathbf{A}_{\mathrm{H}}$ to be an $(M+N)\times(M+N)$ matrix, including also the M queries and assuming that all rows that correspond to queries are 0s. Then let $\mathbf{N}_{\mathrm{H}}$ be the row-stochastic version of $\mathbf{A}_{\mathrm{H}}$, normalizing for dangling nodes—note that all newly introduced queries correspond to dangling nodes—while let $\mathbf{N}_{\mathrm{C}}$ be as before. Finally, let $\mathbf{1} = \mathbf{1}_{\mathrm{C}}$.

For combining the graphs introduce a *querying probability* β, which determines the rate at which a user switches between querying a surfing behavior. The transition-probability matrix for the random walk on the hyperlink-click graph is then given by

$$\mathbf{P}_{\mathrm{HC}} = \alpha\beta\mathbf{N}_{\mathrm{C}} + \alpha(1-\beta)\mathbf{N}_{\mathrm{H}} + (1-\alpha)\mathbf{1}. \qquad (1)$$

Let us also describe at a high-level the random walk defined by the above equation. First, with probability $(1-\alpha)$ the walk goes to a random query or to a random document. With probability α, the walk follows a link in the hyperlink-click graph. The exact action depends on whether the current state is a document or a query. If the current state is a document u, then with probability β the next state is a query q for which there are clicks to u, while with probability $1-\beta$ the next state is a document v pointed by u. If the current state is a query, then with probability β the next state is document for which there are clicks from the query, while with probability $1-\beta$ the next state is any random document.

For our experiments, while we investigate the effects of the value of the parameter β to the results, we fix the value of α to be 0.85, since it is a value widely used for PageRank computation.

5. EXPERIMENTAL EVALUATION

In this section we present the experiments performed in order to validate the utility of the scores produced by random walks on the hyperlink-click graph. We compare these scores to those generated by the hyperlink and click graphs independently. The objective of this section is to discover new information for improving the ranking of Web documents.

For the comparison of the different random walk scores, we focus on two *tasks* in which a good ranking method should perform well. These tasks are: *ranking high-quality documents* and *ranking pairs of documents*. The evaluation is

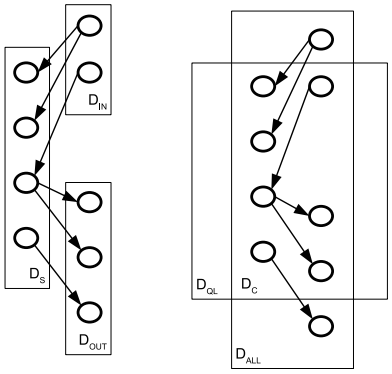

Figure 1: Construction of our dataset.

centered on analyzing the dissimilarities among the different models.

We begin by describing the datasets used.

5.1 Dataset

As a data source we use an in-house query log. Due to the enormous size of the Web, we use only a small sample of documents and queries. Thus, we use only partial graphs instead of the full graphs. No publicly available Web document collections are used, because there are no collections with query log information associated to them, which is a fundamental constraint for our experiments.

We create the graph data by using the query log as the starting point. First let us denote by D_{QL} the set of all documents contained the query log. We parse the query log and we find all the documents that have 10 or more clicks. There are about 9 000 such documents in our sample, and we refer to them as *seed* documents D_{S}.

We then use a Web crawl to find all documents that point to and are pointed to by the seed documents. Let $D_{\mathrm{IN}} = \bigcup_{u \in D_{\mathrm{S}}} N_{\mathrm{IN}}(u)$ and $D_{\mathrm{OUT}} = \bigcup_{u \in D_{\mathrm{S}}} N_{\mathrm{OUT}}(u)$ be the sets of documents with outlinks to and inlinks from D_{S}, respectively, and let $D_{\mathrm{ALL}} = D_{\mathrm{S}} \cup D_{\mathrm{IN}} \cup D_{\mathrm{OUT}}$ be the set of all documents encountered. The above expansion process increases the number of total documents (documents in D_{ALL}) to approximately 144 million.

It should be noted that documents gathered through this expansion process might also exist in D_{QL}. We then define D_{C} to be the documents in the intersection of D_{ALL} and D_{QL}, that is $D_{\mathrm{C}} = D_{\mathrm{ALL}} \cap D_{\mathrm{QL}}$.

Finally, the set of queries Q_{C} that we consider are the queries that have at least one click in the set of documents D_{C}. In total, there are about 61 000 such queries. The dataset construction described above is shown in Figure 1. Given the above sets, we then define the three graphs we consider, the hyperlink graph, the click graph, and the hyperlink-click graph as follows:

Hyperlink graph: The nodes of the hyperlink graph G_{H} are all the documents in the set D_{ALL}. The edges are all the induced hyperlinks between this set of documents. We also note here that, due to the popularity of the documents in the seed set, the set D_{IN} is considerably larger than the set D_{OUT}.

Click graph: The nodes of the click graph G_{C} are the documents in D_{C} and the queries in Q_{C}. The edges are induced by the clicks in the query log, and the number of clicks serve also as weights for the edges.

Hyperlink-click graph: The hyperlink-click graph G_{HC} is the union of the hyperlink graph G_{H} and and click graph G_{C}. Thus the document set for the hyperlink-click is again the set D_{ALL}. The weights on the edges of G_{HC} depend on the querying probability β, as in Equation 1. We use 5 values of the parameter β ($\beta = \{0.25, 0.50, 0.75, 0.85, 0.95\}$), and we denote the resulting graph by $G_{\mathrm{HC}}(\beta)$.

The selected dataset reflects a consistent sample of the Web graph, although highly popular documents are chosen as a seed set, this is further expanded to include most of the neighboring documents. This expansion allows to include in the dataset an heterogeneous sample of documents which are connected to the initial set. Additionally, the query log data is processed very quickly using the MG4J[1] and fastutil[2] tools available on-line. This computational cost is almost negligible compared to that of processing the hyperlink graph.

5.2 Random-walk evaluation

As described in Section 5.1, our experimental datasets are partial and they only represent a sample of the whole Web. Hence, to make the obtained results comparable, we analyze only the results for the documents contained in the intersection of the click, hyperlink and combined graphs (which we refer to as D_{C}). However, it is important to note that we use all of the nodes in each graph to compute the random walk results, and not only the ones contained in D_{C}.

We compute $\boldsymbol{\pi}_{\mathrm{H}}$, $\boldsymbol{\pi}_{\mathrm{C}}$ and $\boldsymbol{\pi}_{\mathrm{HC}}$ for the values of $\beta = \{0.25, 0.50, 0.75, 0.85, 0.95\}$. It is important to take into account that even for very large values of β, random walks on G_{HC} are quite different from those on G_{C}. This is due to the high influence of G_{H} on the combined graph and is observed in throughout the evaluation.

Task: ranking high-quality documents

To compare the random walk results, we decided to focus on high-quality Web documents and how they score within the different models. The hypothesis we sustain is that it is desirable for a good model to score high-quality documents above other documents. To measure this, we use documents from the DMOZ document directory.[3] Our working hypothesis is that since DMOZ is editorially maintained, on average, documents in this directory are of higher quality than documents not in the directory. Consequently, we use D_{Z} to denote the set of documents in the evaluation set D_{C} that belong also to the DMOZ directory. Following our working hypothesis, we postulate that the graph that produces the best ranking results is the graph that ranks documents in D_{Z} higher than the rest of the documents in D_{C}.

To quantitatively measure the agreement of the rankings produced from the different graphs with the DMOZ directory, we use two measures:

Π_{Z}: Our first measure is the normalized sum of the $\boldsymbol{\pi}$ scores of D_{Z} documents. This is,

$$\Pi_{\mathrm{Z}} = \Big(\sum_{d \in D_{\mathrm{Z}}} \boldsymbol{\pi}(d)\Big) / \Big(\sum_{d \in D_{\mathrm{C}}} \boldsymbol{\pi}(d)\Big).$$

[1] http://mg4j.dsi.unimi.it
[2] http://fastutil.dsi.unimi.it
[3] http://dmoz.org

Algorithm 1 Micro-evaluation

1. define a set of queries $Q \subset Q_C \in G_C$ that have at least 1 edge to a document in D_Z and 1 edge to a document in $D_C - D_Z$.
   (a) for each $q \in Q$ find all the adjacent documents D_q that belong to D_C.
   (b) compute Π_Z and Γ_Z replacing D_C with the documents in D_q and D_Z with $D_q \cap D_Z$.
2. Compute the average values of Π_Z and Γ_Z.

The intuition of this measure is that we want a large amount of probability mass of the stationary distribution of the random walk to be accumulated with documents in D_Z. Thus the value of the measure should be as high as possible.

Γ_Z: The second measure we use is inspired by the *Goodman-Kruskal Gamma* measure[13], which is a descriptive rank-order correlation statistic, often used in psychology. Given two rankings on a set of items, on which the two rankings disagree on D pairs of items and agree in A pairs, the Γ measure between the rankings is defined to be $\Gamma = (D - A)/(D + A)$. In our case, even though membership in the DMOZ category does not induce a complete ranking, we can still consider a weak ranking in which all documents in DMOZ are ranked before all documents that are not in DMOZ, and the definition of Γ can still be applied: we just do not include pairs of documents that are either both in DMOZ or none in DMOZ. The measure Γ takes values between -1 and 1, where -1 means that the two rankings are completely discordant, while 1 means that the two rankings are concordant. Again the value of the measure should be as high as possible.

We evaluate the proposed measures Π_Z and Γ_Z in two levels of granularity, which are are defined as follows:

Macro-evaluation: This evaluation intends to capture the overall scores of high-quality documents for the complete D_C document set. The quality measures Π_Z and Γ_Z are computed considering all the documents in D_C and D_Z.

Micro-evaluation: This evaluation is performed at *query level*. This means that to compute Π_Z and Γ_Z the sets D_C and D_Z are reduced to only those documents clicked from a particular query. This is repeated for each query in Q_C that has at least one document in DMOZ and at least one document that is not in DMOZ. In the end the results of are averaged over the total number of queries processed. Formally the procedure for the micro-evaluation is defined in Algorithm 1.

The results obtained in the macro and micro evaluations are shown in Table 1 and Table 2, respectively. The macro-evaluation results show that for the Π_Z the best value is obtained for G_H and the worst for G_C. On the other hand, in the Γ_Z the roles are reversed with G_C being the overall graph with less inverted elements and G_H the one with the most number of inverted elements. The results of the G_{HC} follow closely the best performing scores with less than 0.003 difference for Π_Z and 0.085 difference for Γ_Z.

Table 1: Macro-evaluation results

	Π_Z	Γ_Z
G_C	0.275	**0.643**
G_H	**0.600**	0.458
$G_{HC}(0.95)$	**0.597**	**0.558**
$G_{HC}(0.85)$	0.591	0.552
$G_{HC}(0.75)$	0.587	0.551
$G_{HC}(0.50)$	0.580	0.544
$G_{HC}(0.25)$	0.574	0.540

Table 2: Micro-evaluation results

	Π_Z	Γ_Z
G_C	**0.738**	**0.604**
G_H	0.664	0.273
$G_{HC}(0.95)$	**0.752**	**0.563**
$G_{HC}(0.85)$	0.749	0.546
$G_{HC}(0.75)$	0.745	0.534
$G_{HC}(0.50)$	0.738	0.501
$G_{HC}(0.25)$	0.730	0.483

The micro-evaluation results, in Table 2, shows that for the Π_Z metric, G_{HC} obtains the best value followed by G_{HC}. For the Γ_Z metric G_C is the best, and G_H is the worst in both Π_Z and Γ_Z.

These metrics observe the performance of the random walk scores using different perspectives. From our point of view a good scoring method should perform well both at macro and micro level. The results obtained show that the random walk scores on the G_{HC} follow closely the best scores generated by the non-combined graphs.

Task: ranking pairs of documents

In addition to evaluating our rankings using the measures Π_Z and Γ_Z, which are based on the assumption that documents in DMOZ are on average of high quality, we also perform a user study.

We evaluated a set of triples of the form (q, d_1, d_2) where q is a query with at least 10 clicks in total, d_1 and d_2 are two distinct documents returned by the search engine for that query. Also, we limited the evaluation to cases in which the ordering of d_1 and d_2 was different according to at least two scoring methods in $\{ \boldsymbol{\pi}_H, \boldsymbol{\pi}_C, \boldsymbol{\pi}_{HC} \}$. The evaluation interface is shown in Figure 2. Users where presented a randomly selected triple and asked: "*Is one of these pages clearly better for the query q?*". They were also given the option to say that the two documents were about the same, or that they could not be compared.

A group of 13 human assessors participated in the evaluation. A total of 1,710 assessments were collected, from which 515 (32%) expressed preference for one of the two documents. There were 82 cases in which more than one evaluator assessed the same triple and expressed a preference for one of the two documents. In those cases, the agreement among the evaluators was 70%. Still, the assessment process proved to be very difficult since many of the selected pairs of documents have only a marginal difference in their scores.

The results of the user study are shown in Table 4, using again the Γ statistic to measure the agreement between the

Table 3: Top 10 documents for 3 of the random walk scores

G_{H}	G_{C}	$G_{\mathrm{HC}}(85)$
www.mp3lyrics.org	www.yahoo.com	www.gmail.com
www.gratka.pl	cams.com	www.quizilla.com
www.pimpmyspacepages.com	uk.yahoo.com	www.gratka.pl
www.dpreview.com	www.google.com	www.ebay.com.my
www.mtv.com/...	www.theaa.com/...	www.veoh.com
www.ebay.com.my	www.ebay.co.uk	www.livejournal.com
www.veoh.com	www.nationalrail.co.uk	www.google.pl
www.xe.com	www.cineworld.co.uk	spaces.live.com
www.livevideo.com	games.yahoo.com	www.flixster.com
www.music.com	www.streetmap.co.uk	mail.yahoo.co.uk

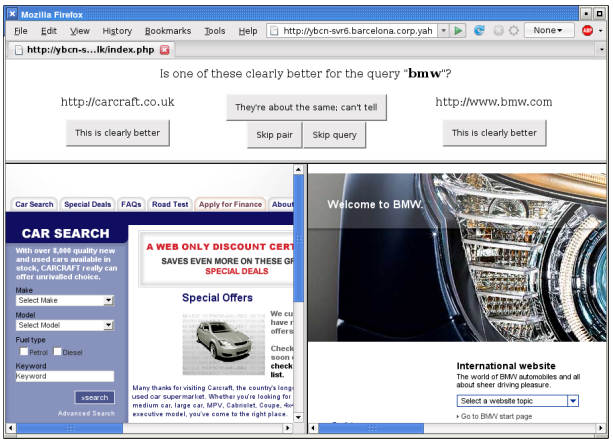

Figure 2: Classification interface.

rankings of the algorithms and the rankings induced by the human evaluators.

Table 4: Γ of ranking functions with human preferences

Method	Overall	Average per query
$\boldsymbol{\pi}_{\mathrm{C}}$	**0.197**	**0.195**
$\boldsymbol{\pi}_{\mathrm{HC}}$	0.063	0.042
$\boldsymbol{\pi}_{\mathrm{H}}$	-0.122	-0.141

Table 5: Γ of ranking functions with human preferences for $\delta \geq 4.5 \cdot 10^{-7}$ (38% of unique pairs)

Method	Overall	Average per query
$\boldsymbol{\pi}_{\mathrm{HC}}$	**0.156**	**0.132**
$\boldsymbol{\pi}_{\mathrm{C}}$	0.111	0.124
$\boldsymbol{\pi}_{\mathrm{H}}$	-0.078	-0.082

Due to the marginal difference in scores between many pairs of documents, we study the behavior of Γ for the pairs of documents which have a *greater* difference between their scores. For this we evaluate only the pairs of documents (d_i, d_j) for which all scoring methods have a minimum $\delta = |\boldsymbol{\pi}(d_i) - \boldsymbol{\pi}(d_j)|$. This allows to evaluate pairs that are less ambiguous to assess for humans. As a result we found that for values of $\delta \geq 4.5 \cdot 10^{-7}$ the Γ values of $\boldsymbol{\pi}_{\mathrm{C}}$ and $\boldsymbol{\pi}_{\mathrm{HC}}$ are reversed and that $\boldsymbol{\pi}_{\mathrm{HC}}$ produces the best performance at this point (shown in Table 5).

If we continue to increase the minimum value of δ we obtain the results shown in Figure 3.

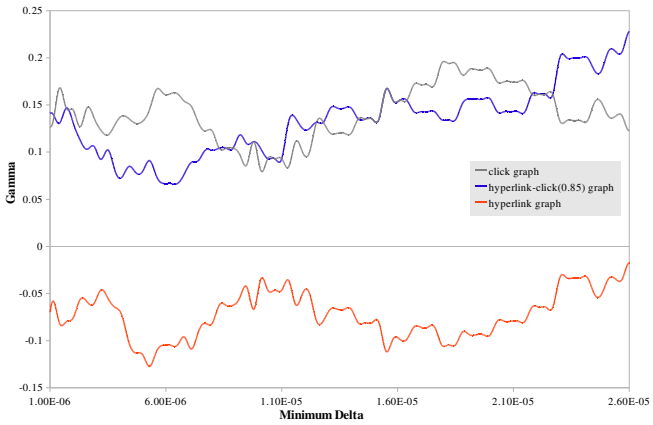

Figure 3: Behavior of Γ in the user study when restricting the minimum allowed value of δ.

Introducing "variations" into the click-graph

The click graph, just as the hyperlink graph can be prone to induced variations, which can affect the scores of the random walk. For the hyperlink graph it is well know that typical variations are produced by link-spam. In the case of the click graph, undesirable modifications in the random walk scores can be the consequence of different methods that increase the number of clicks, such as click-spam. Other variations on the click-through data can occur from sponsored placement of search engine results, in general these do not represent a practical problem, since in general they can be filtered from a query log. Nevertheless we will study the effects of induced variations by using clicks on sponsored results to simulate click-spam.

In the previous part of the evaluation sponsored clicks were filtered from the G_{C} and G_{HC}. We repeat this evaluation introducing sponsored click-through data into G_{C} and G_{HC}. Tables 6 and 7 show the results of the high-quality document evaluation with this variation. We can observe that in the macro-evaluation the order prevails with respect to the original results. On the other hand, in the micro-evaluation G_{HC} performs better for both metrics.

The user study was repeated with 5 judges, which did 1,576 assessments in total, from which 588 (37%) expressed a preference for one of the two documents. Unlike the results

Table 6: Macro-evaluation results

	Π_Z	Γ_Z
G_C	0.2151	**0.7912**
G_H	**0.5851**	0.2103
$G_{HC}(0.95)$	0.5584	0.6429

Table 7: Micro-evaluation results

	Π_Z	Γ_Z
$G_{HC}(0.95)$	**0.5772**	**0.2361**
G_H	0.5713	-0.1495
G_C	0.5356	0.1677

of the user study without sponsored clicks, in this case users agreed more with $\boldsymbol{\pi}_H$ and less $\boldsymbol{\pi}_C$, i.e.: results were reversed. Nevertheless, the results for $\boldsymbol{\pi}_{HC}$ remained in the middle (see Table 8).

Table 8: Γ of ranking functions with human preferences using click-through data with sponsored clicks

Method	Overall	Average per query
$\boldsymbol{\pi}_H$	**0.098**	**0.088**
$\boldsymbol{\pi}_{HC}$	-0.091	-0.137
$\boldsymbol{\pi}_C$	-0.244	-0.186

Summary of the experimental evaluation

In Tables 9 and 10 we provide a concise summary of the metrics and types of evaluations used to measure the quality of the different random walk scores. The convention that we use is that $G_A > G_B$ means that the ranking generated using the graph G_A is better than the ranking generated using the graph G_B (according to our measures), while $G_A \approx G_B$ means that the difference between the two rankings is less than 0.1.

In Figures 4 and 5 we show a comparison of metrics Γ_Z and Π_Z with click variations and without variations. In this Figures we can observe that the values for G_{HC} are always very close or better than the best result from the non-combined graphs. This result is independent on whether or not click variations where induced into the data.

Overall the different tasks evaluated reflect consistency in the results. The values obtained for the study performed with DMOZ documents are coherent for the variations in the value of β, and furthermore, they agree with the results obtained from the user evaluation. We consider this as an indicator of the usefulness of the evaluation and its metrics.

6. CONCLUSIONS

In this paper we studied the effects of a random walk on a unified Web graph. This Web graph combines both hyperlinks between documents and clicks from queries to documents, and was created to capture more completely users' searching and browsing behavior in the Web.

Our main motivation for studying this unified graph is to analyze the new information that it can provide. As a first approach, we focus on the task of using this model to enhance Web document ranking. For this we used a number of different evaluation metrics in order to assess the ranking produced on tour graph with respect to rankings produced by the hyperlink and click graphs. We evaluated by analyzing useful tasks for ranking, such as, ranking high-quality documents and also ranking pairs of documents. For the later, we conducted a user study which provided consistent results with the rest of the evaluation. On the other hand, we also tested the tolerance of our model to click variations or *noisy* data.

Our experimental evaluation shows that the scores generated by random walks on the combined Web graph have several useful properties for document ranking. Overall these scores produce good quality results which are very stable and tolerant to noisy clickthrough data. Additionally, our results show that the unified graph is always close to the best performance of either the click or hyperlink graph. Furthermore the results on the combined graph never approximate the lower bound according to any metric, while the non-combined graphs do not generate good results in all cases.

It is our belief that these properties of the unified graph are useful for improving current ranking techniques. Partly as an indicator of how reliable link-based ranks and click-based ranks are for different tasks. As well as an independent indicator of document quality.

As part of future work we would like to analyze how to deal with the inherent bias that exists in any ranking technique based on usage mining. This is, that pages with already high stationary distribution scores are presented to users more often as a query result. Thus, high ranked pages tend to be more clicked. In long term use, this could create a self-reinforcing ranking. This is not an easy problem to solve and it depends mainly on other underlying ranking techniques. Therefore we recommend any click-based ranking as as a complement to other independent ranking methods.

Also in the future we would like to analyze other Web mining applications for the hyperlink-click graph, such as: link and click spam detection and similarity search.

The code used for the weighted random walk described in this paper is available at `http://law.dsi.unimi.it/satellite-software/`

Acknowledgments. The authors thank Claudio Corsi from the University of Pisa for his help with data preprocessing. Also we thank Debora Donato and Vanessa Murdock from Yahoo! Research for valuable discussions and feedback.

7. REFERENCES

[1] R. Baeza-Yates. Graphs from search engine queries. *SOFSEM 2007: Theory and Practice of Computer Science*, pages 1–8, 2007.

[2] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. Link-based characterization and detection of Web Spam. In *Second International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Seattle, USA, August 2006.

[3] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. pages 407–416, 2000.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, 30(1–7):107–117, 1998.

[5] N. Craswell and M. Szummer. Random walks on the click graph. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 239–246, New York, NY, USA, 2007. ACM Press.

Table 9: Summary of the evaluation for the task of finding high-quality documents

metric	macro		micro	
	without click variations	with click variation	without click variations	with click variations
Γ_Z	$G_C \approx G_{HC} > G_H$	$G_C > G_{HC} > G_H$	$G_C \approx G_{HC} > G_H$	$G_{HC} \approx G_C > G_H$
Π_Z	$G_H \approx G_{HC} > G_C$	$G_H \approx G_{HC} > G_C$	$G_{HC} \approx G_C > G_H$	$G_{HC} \approx G_H \approx G_C$

Table 10: Summary of the evaluation for the task of ranking pairs of documents

metric	without click variation	with click variation
Γ	$G_C > G_{HC} > G_H$	$G_H > G_{HC} > G_C$
$\Gamma(\delta \geq 4.5 \cdot 10^{-7})$	$G_{HC} \approx G_C > G_H$	–

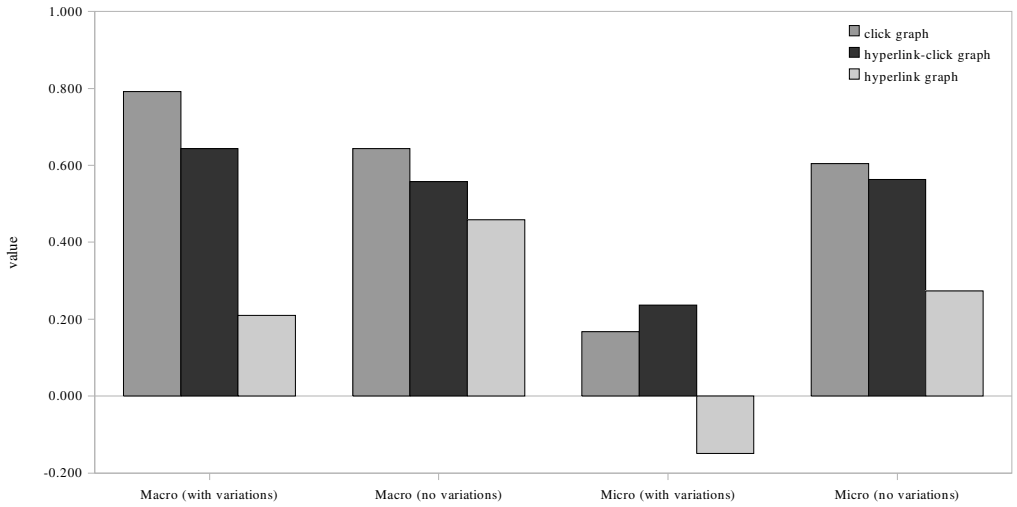

Figure 4: Summary of Γ_Z values with and without click variations for high-quality document evaluation

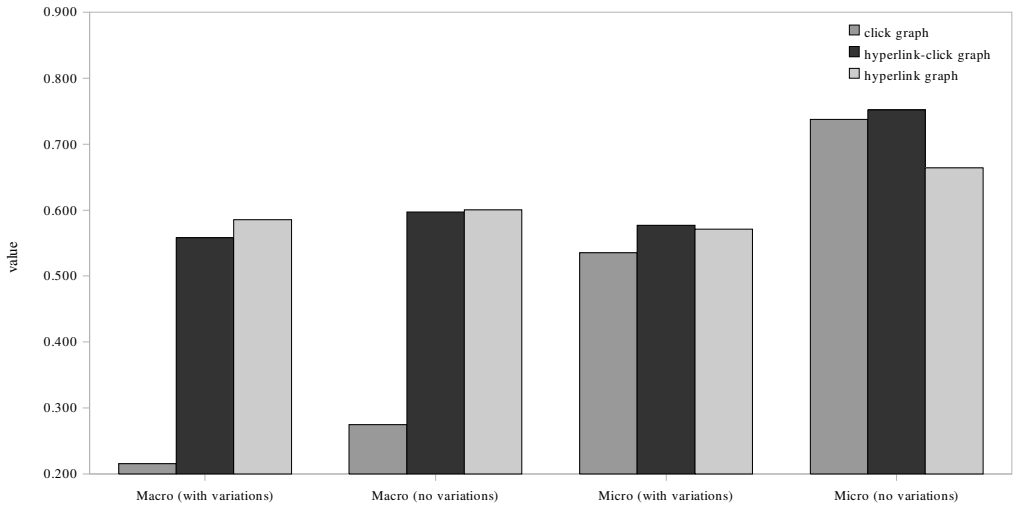

Figure 5: Summary of Π_Z values with and without click variations for high-quality document evaluation

[6] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 87–94, New York, NY, USA, 2008. ACM.

[7] B. D. Davison. Topical locality in the web. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval*, pages 272–279, Athens, Greece, 2000. ACM Press.

[8] G. Dupret, V. Murdock, and B. Piwowarski. Web search engine evaluation using clickthrough data and a user model. In *WWW2007 workshop Query Log Analysis: Social and Technological Challenges*, 2007.

[9] D. Fetterly. Adversarial information retrieval: The manipulation of web content. *ACM Computing Reviews*, July 2007.

[10] Z. Gyöngyi and H. Garcia-Molina. Spam: It's not just for inboxes anymore. *IEEE Computer Magazine*, 38(10):28–34, 2005.

[11] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543, New York, NY, USA, 2002. ACM Press.

[12] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[13] W. Kruskal and L. Goodman. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49, 1954.

[14] M. Lifantsev. Voting model for ranking Web pages. In P. Graham and M. Maheswaran, editors, *Proceedings of the International Conference on Internet Computing*, pages 143–148, Las Vegas, Nevada, USA, June 2000. CSREA Press.

[15] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[16] F. Radlinski. Addressing malicious noise in clickthrough data. In *Learning to Rank for Information Retrieval Workshop at SIGIR 2007*, 2007.

[17] F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 239–248, New York, NY, USA, 2005. ACM Press.

[18] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: discovery and applications of usage patterns from web data. *SIGKDD Explor. Newsl.*, 1(2):12–23, 2000.

[19] W. Xi, B. Zhang, Z. Chen, Y. Lu, S. Yan, W.-Y. Ma, and E. A. Fox. Link fusion: a unified link analysis framework for multi-type interrelated data objects. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 319–327, New York, NY, USA, 2004. ACM.