Where to Start Browsing the Web? *

Dániel Fogaras

[1] Department of Computer Science and Information Theory,
Budapest University of Technology and Economics,
H–1521 Budapest, Hungary
[2] Computer and Automation Research Institute,
Hungarian Academy of Sciences (MTA SZTAKI)
11 Lágymányosi u., H–1111 Budapest, Hungary
fd@cs.bme.hu

Abstract. Both human users and crawlers face the problem of finding good start pages to explore some topic. We show how to assist in qualifying pages as start nodes by link-based ranking algorithms. We introduce a class of hub ranking methods based on counting the short search paths of the Web. Somewhat surprisingly, the Page Rank scores computed on the reversed Web graph turn out to be a special case of our class of rank functions. Besides query based examples, we propose graph based techniques to evaluate the performance of the introduced ranking algorithms. Centrality analysis experiments show that a small portion of Web pages induced by the top ranked pages dominates the Web in the sense that other pages can be accessed from them within a few clicks on the average; furthermore the removal of such nodes destroys the connectivity of the Web graph rapidly. By calculating the dominations and connectivity decay we compare and analyze the proposed ranking algorithms without the need of human interaction solely from the structure of the Web. Apart from ranking algorithms, the existence of central pages is interesting in its own right, providing a deeper insight to the Small World property of the Web graph.

1 Introduction

Recent years witnessed an extensively developing interest on link-analysis algorithms to improve textual based Web search engines. Inevitably, the most influential results on this field are HITS [15,8] and Page Rank [7] algorithms; since then many improvements and extensions appeared [9,17,13], see [5] for a comparative study. HITS assigns a pair of scores to the pages belonging to a query. The authority score of a page is proportional to its importance, and hub score describes the quality of a page as a link collection within the topic. Page Rank, on the other hand, overall quality scores that are applied in any query search later. Following HITS' terminology, the Page Rank scores act as overall authority values of pages independently from any topic. Overall hub scores of the whole Web, however, earned less attention in the link-analysis literature. Remarkable exceptions [18,6] evaluate the rank of a page by summing the rankes

* Research is supported by grants OTKA T 42559 and T 042706 of the Hungarian National Science Fund.

of those linked by the page itself iteratively, which in turn acts as some hub score over the Web.

In the first part of the paper we focus on finding good starting points for browsing from which a large number of pages can be accessed within a few clicks. To express the quality of pages as starting points, we define overall hub scores of the Web, which can be evaluated for the whole Web graph independently from queries. For instance, the hierarchically ordered link collection www.dmoz.org would be given much higher credit as a hub than for example the site www.weather.com with good quality content but only a limited amount of linkage outside its own domain. The Web does not only provide explicitly defined hierarchical link collections that are easy to find, but also contains several implicitly evolving search trees by the nature of hyperlink evolution. The root of such trees are excellent start nodes for browsing, but authority based ranking schemes rarely reveal such root pages.

Clearly it is advantageous to start browsing the Web from a page, if short sequences of clicks from that page lead to as many other pages as possible. We introduce *Start Rank*, a family of hub ranks through counting the search paths departing from each Web page. User defined parameters tune the credit given for each search path. Path-counting method appears first in the classical paper [14] about social networks defining an influence measure, *standing* of persons that is closely related to authority measure of Web pages. We slightly generalize their path-counting technique and apply the method to estimate the hub quality of pages. Notice that hub scores of HITS are proportional to the authority values of directly accessible pages, while Start Rank takes into account the pages accessible in more than one click on hyperlinks.

As a candidate for overall hub ranking, we investigate *Reversed Page Rank* that is computed after reversing the direction of all the hyperlinks, similar ranking algorithms were proposed in [6]. We formally prove that Reversed Page Rank is a member of the family of Start Rank scores supporting the assumption that Reversed Page Rank scores express hub quality. The equivalence of Page Rank and path-counting rank is interesting in its own right stating that Page Rank generalizes in-degree rank by taking into account longer than one-step paths.

Evaluating and comparing the performance of link-analysis algorithm seems hard, since there is no formal definition for the "qualities" of a Web pages. Typical practical approaches are based on expert evaluation [2], volunteer testing [21], notions of "spam" [10] or query examples [5], all depend on human judgment. In a theoretical approach one can formally analyze certain desirable features of ranking algorithms such as stability [3,19], locality and monotonicity [5]. These features are natural requirements for ranking algorithms, but neither of them acts as an objective measure of the quality of link-analysis algorithms.

We propose *centrality analysis* as a graph based tool to provide quantitative justification and comparison of the introduced ranking methods. The key idea is that top ranked pages play a central role in maintaining the connectivity of the Web graph. For any ranking over the nodes of the Web graph, the centrality of the set of top ranked pages can be evaluated numerically, yielding a qualification of the ranking algorithm. Although such a qualification only classifies the top scores assigned for the pages, we believe that centrality analysis is an important step towards the automatic evaluation of ranking algorithms.

The centrality of a set of pages is either measured by *domination*, the average distance from the set to the other pages; or by the decay in the *diameter* of the Web graph after the removal of the central nodes. The former centrality measure is applied for hub ranking schemes, since from the set of strongest hubs the whole Web should certainly be available within a few clicks on the average. The latter notion of centrality will show the quality of a ranking algorithm that gives credit for popular hubs—nodes that are contained in a large amount of search paths of the Web. Our notions of centrality were motivated by the NP-hard combinatorial optimization problem k-domination [11] and by the experiments of [1] measuring the failure tolerance of real networks against removing the largest degree nodes.

Besides qualifying the outputs of ranking algorithms, centrality analysis also provides a deeper insight to the *small world phenomenon*, empirically proved for many implicitly evolving networks including the Web graph [4]. A network is referred to as a small world, if the diameter is low and the number of edges in the network is relatively small. Centrality analysis experiments reveal that a surprisingly small number of central nodes are responsible for the connectivity of the Web graph. Such experiments were pioneered by [1] implying that only a small set of largest degree nodes maintains the connectivity of small world networks. In our centrality analysis experiments we strengthen the results of [1] by showing more centralized nodes than the pages with largest degrees.

Our experiments containing both centrality analysis and query search examples were conducted on the `.ie` domain, the Irish Web. While this portion of the Web provides a computationally feasible test-bed, the contextual structure of a national domain will not differ so much from the entire Web that would result in significant bias in the experiment. Ranking is performed over the collection of near one million pages crawled in October 2002; however for keyword searches we also relied on queries to Google [12].

2 Hub scores of the Web

Finding a good start page is a critical part of browsing the Web: it is clearly worth starting from a site from which a large amount of content can be reached within a few clicks. Slightly modifying the notion of Kleinberg's HITS algorithm [15] we refer to such pages as *good hubs*.

In this section we introduce Start Rank as a family of hub scores to measure the quality of pages as start nodes. Then we show one member of this family easily computable by slight modifying Page Rank. Finally, some combinations with other ranking algorithms are proposed.

2.1 Start Rank

Start Rank assigns a hub score for each web page on the basis of counting the search paths originating from the page in question. Each search path is taken into account by a weight depending on the value of the target page, the length of the search path and the hyperlinks occurring in the search path. Note that Start Rank naturally generalizes out-degree as the simplest measure on the hub quality of pages, since out-degree counts all the one-step walks from each page.

The actual Start Rank scores are determined solely from the structure of the Web graph and from the following three user defined parameters.

- The *length weight function* assigns a real weight $\ell(i) \geq 0$ for each length $i \geq 0$. The requirement that longer search paths generally worth less than shorter paths can be achieved for a Start Rank by setting a monotone decreasing length weight function. Furthermore, to eliminate the false effect of extremely long search paths containing a large amount of cycles, it is reasonable to choose zero length weight beyond a threshold. In most of what follows exponentially vanishing length functions are employed with expected value falling into the range 5-15.
- The *target value* $t(v) \geq 0$ of a page v emphasizes the credit that is given for a search path for finding v. Setting the target value identical over the Web pages implies that all pages are treated equally worth as targets. Alternatively, an overall quality measure, such as the Page Rank [7] can be chosen as target value for each page. Then a node obtains high Start Rank, if a large amount of search paths lead to high quality pages from the node in question. Another approach is to set the target value topic specific by giving positive value only for a collection of pages inducing a topic of the Web.
- The *link factor* $m(u \to v)$ assigns a real weight for each hyperlink $u \to v$ of the Web. The appropriate choice of $m(u \to v)$ is inversely proportional to the effort spent by a surfer to select the link from the page u, when proceeding in a search path. For example, the effort can be measured by $d^+(u)$, the number of out-going links from page u, thus $m(u \to v) = \frac{1}{d^+(u)}$ can act as a link factor. More intimate link factor settings take into account the position or size of the anchor text of the hyperlink in the HTML document.

Definition 1 *For given user defined parameters –length weight function, target values, and link factors– the* weight $w(P)$ of a search path P *with length i, target node v is defined as follows,*

$$w(P) = t(v) \cdot \ell(i) \cdot \prod_{e \in P} m(e),$$

where the product is taken over all link e contained by the path. The start rank $SR(u)$ *of a node u is*

$$SR(u) = \sum_{P:\ u \rightsquigarrow v} w(P),$$

summing over all paths originating at u.

In the rest of this section, we show that the n-dimensional *Start Rank vector* $\underline{\mathrm{SR}}$ can be expressed as a linear combination of matrix powers, where n denotes the number of Web pages. Let M denote the n-by-n matrix with entries $M_{v,u} = m(u \to v)$ for each link $u \to v$; and $M_{v,u} = 0$, if the link $u \to v$ does not exist. (Equivalently, M is obtained by transposing the adjacency matrix of the Web graph and by replacing each 1 entry with the link factor corresponding to the directed edge.) Furthermore, by introducing the $\underline{t}$ notation for the n-dimensional row vector of the target values, the weights arising for search paths with length i are $\ell(i) \cdot \underline{t} \cdot M^{i}$, thus the start rank scores can be expressed as

$$\underline{\mathrm{SR}} = \sum_{i=0}^{\infty} \ell(i) \cdot \underline{t} \cdot M^{i}. \tag{*}$$

Evaluating such a formula seems hopeless due to the huge dimensions of M, however, the complexity of multiplying a vector with M is proportional to the number of non-zero entries of M, or equivalently the number of hyperlinks of the Web. Such a multiplication can be performed by external memory implementation, similarly to a Page Rank iteration [7]. Thus, if the length function vanishes for numbers over k, then the $\underline{t}M^{i}$ vectors can be evaluated with k external memory iterations even for the whole Web graph.

2.2 Reverse Page Rank

Since Page Rank (PR) acts as a successful authority score over the Web pages, one may intuitively feel by symmetry that reversing the direction of the hyperlinks and then applying PR yields an overall hub score of the pages. To justify the statement we formally prove the equivalence of reverse Page Rank with a special case of Start Rank scores with appropriate user parameter settings.

For the sake of simplicity in the rest of the paper, we assume that nodes with zero in- or out-degrees have been removed from the Web graph. Furthermore, *reversed Web graph* refers to the graph obtained from the Web graph by reversing the directions of the edges.

First, we recall the definition of PR scores defined on the Web graph through the *random surfer model* resembling the behavior of human users. The surfer takes a random walk visiting the Web sites by selecting the next page according to the following rule: with probability $1-d$, the next page is chosen from those pointed by the currently visited page; and with probability d, it is selected from all the pages according to some *jump distribution* independently from the currently visited page. Intuitively, the above *damping factor* d is the probability that the random surfer gets bored and restarts surfing; in practical applications it is set to $d \approx 0.1 - 0.2$. The jump probabilities describes the preference of the random surfer among starting nodes to jump; in the simplest case this is uniform over all the Web pages. The random surfer model yields a Markov chain and the PR of a Web site is defined as the probability of the page in its stationary distribution [7].

Definition 2 *For given damping factor and jump probabilities, the* reverse Page Rank (RPR) *is defined as the PR computed on the reversed Web graph.*

Similar to Page Rank implementation [7], RPR can be computed by the power iteration method, and it can be evaluated for such an enormous input as the Web graph. RPR can be easily interpreted in the random surfer model, with the modification that the random surfer follows the links backwards. However, the interpretation does not support the assumption that RPR is useful as a hub score—in the rest of this section we deduce that RPR is a member of the family of start rank (SR) scores.

Theorem 1. *The RPR with damping factor $0 < d < 1$ and given jump probabilities is equivalent to a SR with the following parameter settings. The length weight $\ell(i) = d \cdot (1-d)^i$, the target values are identical to the jump probabilities and the link factor $m(u \to v) = \frac{1}{d^-(v)}$ is inversely proportional to the in-degree of v.*

Proof. Let $\underline{j}$ denote the n-dimensional row vector of the jump probabilities and J the $n \times n$ matrix with all rows equal to $\underline{j}$ where n is the number of web pages. Let $\underline{\text{RPR}}$ and $\underline{\text{SR}}$ denote the RPR and SR vectors. Furthermore the stochastic matrix M is obtained from the adjacency matrix of the reversed Web graph by normalizing its rows. Note that normalization is equivalent to multiplying the entries of the adjacency matrix with the corresponding link factors.

For the transition matrix Π of the Markov chain defined by the random surfer model the following equation holds,

$$\Pi = dJ + (1-d)M.$$

Since RPR is the stationary distribution,

$$\underline{\text{RPR}}\,\Pi = \underline{\text{RPR}}. \tag{**}$$

In order to show that the equation $\underline{\text{RPR}} = \underline{\text{SR}}$ holds, we will prove that $\underline{\text{SR}}$ satisfies (**). The $\underline{\text{SR}}$ probabilities can be expressed by equation (*),

$$\underline{\text{SR}} = \sum_{i=0}^{\infty} d(1-d)^i \underline{j} M^i,$$

since the length distribution is geometric with parameter d. By substituting this into (**)

$$\begin{aligned}
\underline{\text{SR}}\,\Pi &= \underline{j}\Big(\sum_{i=0}^{\infty} d(1-d)^i M^i\Big)(dJ + (1-d)M) \\
&= d\,\underline{j}J + \underline{j}\sum_{i=1}^{\infty} d(1-d)^i M^i \\
&= d\,\underline{j} + \underline{j}\sum_{i=1}^{\infty} d(1-d)^i M^i \\
&= \underline{j}\sum_{i=0}^{\infty} d(1-d)^i M^i \\
&= \underline{\text{SR}}.
\end{aligned}$$

The second equation comes from the fact that the matrix $N = \sum_{i=0}^{\infty} d(1-d)^i M^i$ is stochastic, and $NJ = J$ holds for any stochastic matrix, as the rows of J are equal. Similarly $\underline{j}J = \underline{j}$ was applied for the third equation.

Finally, we mention that a similar statement holds for the original PR citation index showing that the PR of each page can be expressed as the weighted sum of all paths arriving at the node in question. Hence PR generalizes the simple in-degree rank by taking into account all the in-coming walks not only the one-step paths.

2.3 Mixed and aggregate ranks

We investigate the alternatives to combine Reverse Page Rank (RPR) with other ranking strategies to obtain refined quality measures on Web Pages. From the several possible options, we especially focus on combinations with ordinary Page Rank (PR) — for more general aggregating methods we refer to [10].

The RPR of each page counts the short search paths leaving from the actual page, and the credit given for a target page can be tuned by setting the target value or equivalently the jump probability of the target as stated in Theorem 1. We propose the following methods for tuning the jump probabilities (target values) of RPR.

- *Uniform RPR* algorithm performs iterations with uniform jump distribution over the Web pages. Such a choice of jump probabilities raises the hub score of pages from which a large amount of nodes can be accessed, however the qualities of the accessed pages are not taken into account. In what follows, we always refer to uniform RPR, if the jump probabilities are not defined explicitly.
- *Popular RPR* algorithm precomputes ordinary PR, and then performs RPR iterations, where the jump probabilities of the nodes are set to the precomputed PR scores. By the assumption that ordinary PR measures the quality of pages, popular RPR will be raised for those pages from which a large amount of high quality content can be accessed within short click streams. Notice the analogy with HITS algorithm [15], where the hub score of a node is equal to the sum of the authority scores available with one step. Popular RPR refines this idea by taking into account the authority scores of nodes available in more than one step with exponential decreasing relevance in the number of clicks.
- *Personalized RPR* assigns non-zero jump probabilities only for the members of a certain topic of the Web following the idea of [20] originally proposed for PR. Personalized RPR scores then express hub quality only in a certain topic. Such approach seems practical for query searches or clustering, while personalized RPR would require on-line computation over the entire Web graph for each topic query.
- *Topic sensitive RPR* acts as an off-line alternative of personalized RPR by computing RPR with a few topic specific jump distributions belonging to some low-dimensional basis of the topic-space. Then, hub scores of an arbitrary topic are evaluated as some linear combination of the basis hub scores, which is practically computable on-line. The method was introduced in [13] for PR and the adaptation is straightforward for RPR.

Fixing the jump distribution with one of the above methods RPR algorithm yields scores expressing the quality of pages as hubs. Such score may present as a component of some overall quality measure of pages as in the following examples.

– *Mixed PR* refers to the family of scores evaluated as $f(\mathrm{PR}, \mathrm{RPR})$, i.e., some function of the already computed PR and RPR values. Mixed PR is a trade-off between hub and authority scores depending on function f.
– *Product PR* score of each page is defined as the product of PR and RPR values. (Notice that product PR specializes mixed PR.) Web pages possessing high product PR are both valuable hubs and authorities, so the numbers of in-coming and out-going paths are both large. We believe that such pages play an important role in maintaining the connectivity of the Web graph.

3 Centrality analysis

For a given ranking of the Web pages, centrality-analysis experiments numerically evaluate the centralities of small sets of top-ranked pages in the Web graph. Such an experiment requires graph theoretical definition of centrality; in the following section we propose different notions of centrality based on averaging some distances in the Web graph.

Distance averaging techniques face the problem of infinite distances that is handled by harmonic mean in our definitions. A further advantage of harmonic mean that it expresses the expected search efficiency of a surfer following the shortest paths of the Web.

3.1 Domination of a start set

From a general start set of pages most other nodes of the Web graph should be available within a few clicks. We introduce a qualification for start sets and an intuitive explanation of the formula through search efficiency.

Suppose that a user is searching for some target page. Let us assume that by carefully reading the contents of the intermediate pages, it is always possible to choose the best possible direction towards the target. In this case the surfer will follow a shortest path.

Next we consider how efficiently the user spent browsing time to find the target. If the target is reached in 3 clicks for example, then he spends one third of his time to read something interesting while the rest of it is wasted for visiting inner pages of the search path. Hence we say that the *efficiency of a start page* s to find target t is $\frac{1}{\mathrm{dist}(s,t)}$, where $\mathrm{dist}(s,t)$ denotes the minimum number of clicks to reach t from s. If there is no path from s to t, then $\mathrm{dist}(s,t) = \infty$ and the efficiency is zero.

More generally, the surfer uses some start set V_S of pages to find target t. As he always starts from the members of V_S, he knows well the contents of these pages. Therefore he can guess the closest page of V_S to t. Then the *efficiency of the start set* is $\frac{1}{\mathrm{dist}(V_S,t)}$, where $\mathrm{dist}(V_S,t)$ denotes the minimum of distances from the nodes of V_S to t. The *domination of a start set* is defined as the average efficiency over all possible web pages as goals. This can be interpreted as the expected efficiency, if a surfer starts searching a random goal page. Formally, the domination of V_S is determined as follows:

$$\mathrm{dom}(V_S) = \frac{1}{|V| - |V_S|} \sum_{t \in V \setminus V_S} \frac{1}{\mathrm{dist}(V_S, t)},$$

where V denotes the set of Web pages. Thus the domination of a start set is the inverse of the harmonic mean of distances between V_S and all the other Web sites.

Our first notion of centrality of a set of pages is equal to the above introduced domination. In the centrality analysis experiments of Section 4.2 we successively add the top ranked pages to a start set and evaluate the domination in each iteration. The experiment reveals the quality of the ranking algorithm to select graph theoretically good sets of hubs or starting points from which the rest of the Web is accessible within a few clicks on the average.

Our notion of domination resembles of the NP-hard combinatorial optimization problem of finding a minimum size subset of nodes in a graph G such that all the other nodes are within a given distance k from the subset [11]. In our scenario such a subset would be a start set from which the farthest node has distance at most k. Such a worst-case analysis cannot express a fine quality measure on the start set, hence we proposed to take the average of distances.

3.2 Attacking the Web

Besides domination, the centrality of a set of nodes can be measured by the *attacking ability* of the set—the decay in the connectivity of the Web graph after removing the set of nodes in question. In our centrality analysis experiments, the top ranked nodes are removed gradually, and then we evaluate the connectivity of the remaining part of the Web graph.

The connectivity is expressed by the *harmonic diameter* of the Web graph, the harmonic mean of distances between all the pairs of nodes. The reciprocal of the harmonic diameter, under the notion of the previous subsection, means the expected efficiency when a surfer starts searching a random goal from a random start node. Hence what we actually measure is the fraction of time spent on reading topics of interest in contrast to downloading pages just to find an appropriate link to move on. Formally if V denotes the set of Web pages, then let

$$\mathrm{diam} = \frac{|V|(|V| - 1)}{\sum_{u \neq v \in V} \frac{1}{\mathrm{dist}(u,v)}}.$$

Another advantage of our notion of harmonic diameter compared to other notions of diameter is that pairs of nodes unreachable from one another have contribution zero in the formula, hence harmonic distance measures both distance and reachable at the same time.

The idea of removing some small portion of Web pages and measuring how the diameter increases was originally proposed [1] for different purpose. They concluded that the failure caused by randomly chosen nodes hardly effect the connectivity of the Web, but an intentional attack removing the nodes with large degree raises the average distance rapidly. Notice that the degrees of nodes also induce a ranking on the nodes. In our experiments we investigate the effect of replacing degree rank with more subtle scores of the importance of pages.

4 Experimental results

Our experiments were conducted on the `.ie` domain, the Web pages of Ireland. We believe that the structure and diversity of this domain is similar to that of the whole WWW. The graph of the `.ie` domain was small enough to store in internal memory, thus any variant of the proposed ranking algorithms were calculated within 15 minutes.

We downloaded 986,207 pages from the Irish Web in October, 2002. We used the open source Web robot Larbin [16] on a 1.8GHz Pentium IV CPU with a 10Mb Ethernet connection. The Web graph induced by the `.ie` domain had 792,902 nodes [3] and 10,037,951 edges. The ranks PR, RPR, popular RPR and product PR were computed with damping factor $d = 0.2$ using 100 power iterations that yielding an error smaller than 10^{-8} in all cases.

4.1 Ranking keyword search hits

We investigate how well RPR or popular RPR serve in ranking keyword queries. We believe that by the nature of ranking link collections high our ranking strategies act well for a broad topic search—at least as a possible aggregated rank component combined with text and link based strategies. In our experiment we submitted keywords of broad topics to Google [12] and saved all the enumerated URLs. The number of available URLs was varying between 500 and 1000. Then we used RPR to reorder these URLs and compared the top ten Google hits with our ranking. Since the reordered list was computed from Google's top $500 - 1000$ hits, this can be treated as an aggregate of Google's ranking with popular RPR.

The query results are listed on Table 4.1 for "fishing" and "sailing"—typical broad topic query strings for exploring certain topic rather than searching for a specific piece of information. The number 1, 4 and 5 hits of Google on "fishing" are Web sites of specific famous fishing resorts and boats—inevitably these pages provide popular content. Popularity is however not appreciated by the RPR scores; instead credit is given to link collections. Such examples are the number 2, 4, 5, 7 and 8 hits of RPR query or 1, 2, 3 and 4 of popular RPR for "fishing". Hit number 8 of popular RPR on "sailing" is a remarkable example of a good link collection. Such a collection may act as an excellent start node to explore "sailing in Ireland".

A drawback of RPR and popular RPR can be also read from the lists of top ranked URLs. Both gives high credit to archives or large collections of databases within a Web site. Such examples are 1 and 3 from RPR with query "fishing". In some cases popular RPR was able to overcome the problem such as in the case of "fishing" query, since the members of the archive have low target probability.

[3] We have deleted those pages not linking within the `.ie` domain that would otherwise correspond to a node with zero out-degree in the graph.

Table 1. Query results for Google and by reordering the top 500-1000 hits of Google.

Google with query “fishing” 1 indigo.ie/~bwlodge/ 2 indigo.ie/~bwlodge/fisreport.htm 3 www.infowing.ie/fishing/ 4 www.infowing.ie/fishing/Sligo2.htm 5 homepage.tinet.ie/~bluewater/ 6 homepage.tinet.ie/~ncffi/ 7 www.shannon-fishery-board.ie/ 8 www.shannon-fishery-board.ie/ fishing-open.htm 9 www.react.ie/Activities/Fishing.htm 10 www.react.ie/Activities/ Fishingwhere.htm	Google with query “sailing” 1 www.sailing.ie/ 2 www.iol.ie/ glenans/ 3 www.iol.ie/ gerbyrne/ 4 www.braysailingclub.ie/ 5 www.braysailingclub.ie/sailing/ sailing_instructions.html 6 www.alia.ie/sailing/ 7 www.alia.ie/sailing/afloat.html 8 www.arklowsc.ie/ 9 www.arklowsc.ie/Sailing_Tips/ sailing_tips.htm 10 homepage.tinet.ie/ bmcg/Cullaun/ cullaun.htm
RPR with query “fishing” 1 www.ndpgenderequality.ie/statdata/ 2002/measure/measure4.html 2 www.nci.ie/holiday 3 www.ndpgenderequality.ie/statdata/ 2002/topic/topics17.html 4 kildare.local.ie/things_to_do_and_see 5 www.lakedistrict.ie/fishing/index.shtml 6 www.thecia.ie/patricks 7 westmeath.local.ie/things_to_do_and_see 8 www.oksports.ie/irish/water.html 9 www.falconholidays.ie/locations/ 12/11.html 10 www.cybercottage.ie	RPR with query “sailing” 1 sport.startpage.ie 2 www.irishferries.ie/sitemap.shtml 3 www.homefromhome.ie/properties.asp 4 www.kellyco.ie/html/AvailRes.html 5 www.athlonechamber.ie/about-athlone/ tourism.htm 6 www.oksports.ie/irish/water.html 7 www.wolfhound.ie/eveningclasses/ email.htm 8 doon.mayo-ireland.ie/moores.html 9 www.inside.ie/e_article000074755.cfm 10 www.csis.ul.ie/staff/CiaranCasey/ personal.htm
Popular RPR with query “fishing” 1 www.nci.ie/holiday 2 kildare.local.ie/things_to_do_and_see 3 www.infowing.ie/fishing 4 www.lakedistrict.ie/fishing/index.shtml 5 www.connacommunitycouncil.ie 6 westmeath.local.ie/things_to_do_and_see 7 www.thecia.ie/patricks 8 tiara.ie/goingto.htm 9 indigo.ie/~bwlodge/fisreport.htm 10 www.cybercottage.ie	Popular RPR with query “sailing” 1 www.irishferries.ie/sitemap.shtml 2 sport.startpage.ie 3 www.kellyco.ie/html/AvailRes.html 4 www.homefromhome.ie/properties.asp 5 www.athlonechamber.ie/about-athlone/ tourism.htm 6 www.wolfhound.ie/eveningclasses/ email.htm 7 www.rte.ie/aertel/p581.htm 8 www.oksports.ie/irish/water.html 9 www.tourismresources.ie/fh/ shannon.htm 10 www.rosscarbery.ie

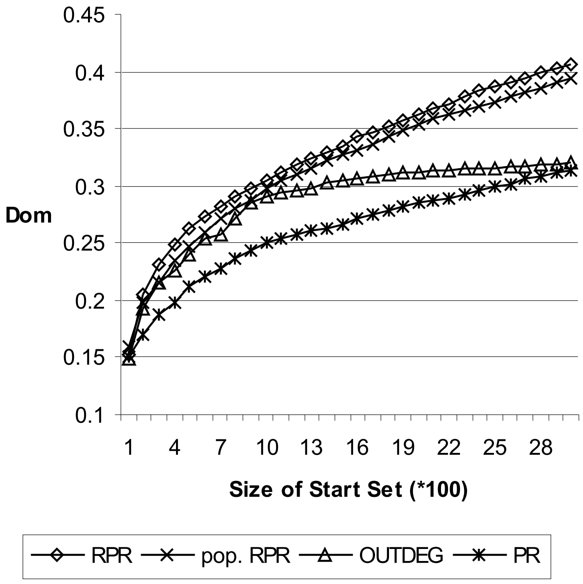

Fig. 1. Domination of start sets.

4.2 Top ranked pages, domination and diameter

In our centrality analysis experiments we selected the first few top ranked pages under different ranks and measured graph theoretic quantities related to distance and connectivity as a function of the number of pages selected. We graphed our results for multiples of a hundred Web pages. Under any reasonable ranking strategy the top few hundred nodes should form a subset of the Web with an important role in search and navigation. Note that the size in question is smaller than one percent of our document collection of pages from Ireland.

In our first experiment we constructed start sets of sizes falling into the range $100-3000$ from the top ranked nodes. Then we calculated the domination of these sets, the efficiency of searching a random node from the given set. The results for PR, RPR, popular RPR and out-degree rank are depicted on Fig. 1.

The diagram shows that nodes with large PR behave worse as start sets than even the simple heuristic of choosing out-degree as rank. On the other hand both RPR and popular RPR finds sets with large domination, i.e., from these sets all the other pages are accessible within a few clicks on the average. Recall that RPR scores are based on counting the weighted sum of all search paths as stated in Theorem 1. The domination of top ranked sets are calculated on the basis of shortest paths, thus we conclude from the success of RPR scores that RPR acts as some approximation of shortest path counting. We mention that such approximation results do not hold in arbitrary graphs, since in RPR all the search paths are taken into account not only the shortest paths.

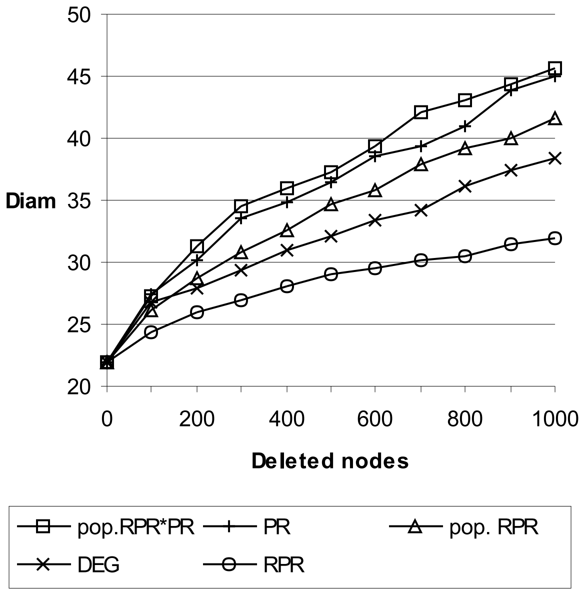

Fig. 2. Increasing the diameter of the Web graph by removing the top ranked nodes.

The removal of the top ranked nodes should, in addition to having large domination, also destroy the connectivity of the Web. While removing the top ranked 100, 200, ..., 1000 nodes, we measured the harmonic diameter of the remaining graph. [4] The results are depicted on Fig. 2 for PR, RPR, popular RPR, degree rank, and the mixed rank computed as the product of PR and popular RPR.

Product PR turns out the strongest "destructor" by increasing the diameter over 45 after removing 1000 nodes. The reason for this phenomenon is that product PR can only be high for a node having both high RPR and PR scores. High RPR scores imply that a large number of search paths depart from the page, and the PR score shows that a large amount of search paths arrive at the node in question. Thus, a node with high product PR is a typical inner node of short search paths of the Web. Therefore the removal of such central nodes destroys the connectivity of the Web as verified by our experimental results.

The fact that RPR has the lowest power of destruction among the measures appears surprising and contradicting the domination results. However it is easy to put the two results together and conclude that top start rank nodes, instead of acting central and interconnecting different topics and domains, serve for finding quick routes by possibly sitting on the top of large semi-local collections of specific and non-overlapping topics.

Except for RPR, all the ranking algorithms performed better than the degree rank, thus we strengthen the results of [1]. PR, product PR and popular RPR all provide

[4] An exact computation of the diameter would require a Depth First Search from each node. Thus we approximated the result by computing DFS from 1000 randomly chosen nodes.

central sets of nodes taking the responsibility for the low diameter of the Web graph. The existence of such centralized sets let us a deeper insight how the small world property is achieved for the Web graph.

5 Conclusion

Start nodes play important roles in exploring some part of the Web. We proposed start rank algorithms to express the qualities of pages as hubs based on short random walk arrival probabilities. The algorithm performs Page Rank computation on the reversed Web Graph. Thus, it is practically implementable in case of the Web graph. Graph theoretical tools are introduced to evaluate start ranking algorithms by measuring the domination and the attacking ability of the top ranked nodes. In our experiments on the Irish Web, the proposed start ranking algorithms selected start sets with largest domination justifying our intuitions. We believe that aggregating the start rank algorithms in text based query search engines improves the efficiency of browsing the Web.

6 Acknowledgment

I wish to thank Katalin Friedl, András Benczúr and András Lőrincz for the valuable discussions and for improving the level of this manuscript.

References

1. R. Albert, H. Jeong, and A. Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000.
2. B. Amento, L. Terveen, and W. Hill. Does authority mean quality? Predicting expert quality ratings of web documents. In *Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2000.
3. Y. Azar, A. Fiat, A. R. Karlin, F. McSherry, and J. Saia. Spectral analysis of data. In *ACM Symposium on Theory of Computing*, pages 619–626, 2001.
4. A.-L. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: the topology of the word-wide web. *Physica A*, 281:69–77, 2000.
5. A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Finding authorities and hubs from link structures on the world wide web. In *10th International World Wide Web Conference*, pages 415–429, 2001.
6. J. Boyan, D. Freitag, and T. Joachims. A machine learning architecture for optimizing web search engines. In *Proceedings of the AAAI Workshop on Internet-Based Information Systems*, 1996.
7. 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.
8. S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the Web's link structure. *Computer*, 32(8):60–67, 1999.
9. B. D. Davison, A. Gerasoulis, K. Kleisouris, Y. Lu, H. ju Seo, W. Wang, and B. Wu. Discoweb: Applying link analysis to web search. In *Proceedings of the 8th World Wide Web Conference, Toronto, Canada*, 1999.

10. C. Dwork, S. R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *10th International World Wide Web Conference*, pages 613–622, Hong Kong, 2001.
11. M. Garey and D. Johnson. *Computer and Intractability : A Guide to the Theory of NP-completeness*. W.H. Freeman, San Fransisco, 1979.
12. Google. Commercial search engine founded by the originators of pagerank. located at `http://www.google.com`.
13. T. H. Haveliwala. Topic-sensitive pagerank. In *11th International World Wide Web Conference*, Honolulu, Hawaii, 2002.
14. L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, March 1953.
15. J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
16. Larbin. Multi-purpose web crawler.
17. R. Lempel and S. Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. In *9th International World Wide Web Conference*, 2000.
18. M. Marchiori. The quest for correct information on the web: Hyper search engines. In *7th International World Wide Web Conference*, 1998.
19. A. Y. Ng, A. X. Zheng, and M. Jordan. Stable algorithms for link analysis. In *Proc. 24th Annual Intl. ACM SIGIR Conference*, 2001.
20. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
21. M. Richardson and P. Domingos. The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. In *Advances in Neural Information Processing Systems 14*. MIT Press, 2002.