

An Approach to Confidence Based Page Ranking for User Oriented Web Search

Debajyoti Mukhopadhyay, Debasis Giri, Sanasam Ranbir Singh

Department of Computer Science & Engineering
Haldia Institute of Technology

P.O. Hatiberia, Haldia, Dist. East Midnapore, WB 721657, India
(debm@vsnl.com, debasis-giri@mailcity.com, san_ranbir@yahoo.com)

1. Introduction

Searching the World Wide Web considered as a digital library having widely distributed information for news, advertisements, education, government, e-commerce and many information poses a great challenge. Keyword based search engines like Lycos [12], Excite [15], Northern Light [16], Alta Vista [14] often suffers from lack of quality pages. Information on the Web is authored by millions of Web creators having a variety of backgrounds, knowledge and interests. Web can be viewed as a graph, where pages are the nodes and edges are the hyperlinks. In recent studies, hyperlinks are being actively used to improve Web search performance [2, 5, 11, 8, 12]. Hyperlinks are created by web creators based on their interest on the cited pages and/or relevance to the cited pages with their pages. Although, hyperlink (or citation) based search engines like Google [1] provides better search quality, communities [3] and confidence of the pages for the search topic cannot be left out. Searching for related pages on this vast library is a tricky job and finding the most relevant pages for specific topic from the related pages poses the same degree of complexity. Basic intuition says that a customized graph will enhance and provide better search result.

The main goal of this paper is to customize the Web for specific feature and/or community graph, finding out the confidence of each page in the graph in question from the past experience and calculate the page rank of the pages in the graph from confidence obtained and link structure. We view the Web in the Universe from the users query points of view and customize accordingly.

Several studies find that Web self-organizes such that communities of the highly related pages can be efficiently identified. Identifying the communities on the Web can enhance focused search engines. A Web community is defined as a collection of Web pages such that each member

page has more hyperlinks in either direction within the community than outside the community [5]. Related pages are often linked to each other. Pages in a community can be assumed as the focused region for searching significant pages related to that community. Community plays a significant role in customizing the Web.

Link structure analysis can improve search engine [8]. Basic assumption in citations based Web search is that a citation is made because of the Web creators' interest on the cited page and considered to be related to each other. From several studies [4, 6, 7, 9, 10] it is found to be true in most of the cases. But, cited page and citing page may not always necessarily have similar features. For example, personal Web pages often have links to the different pages of different categories like hobbies, institutions, favorites etc. For search like institute, it is meaningless to include the home page in question into the page search region. In our study, we removed all such pages and citations from such pages are not taken into consideration for page rank calculation. But, such pages cannot be ignored for extracting features and/or categories. Using extended anchor-text and full-text, Web pages can be effectively classified [3].

In most search engines, page maintained in different URLs have often found to be listed more than one time in the top list, suppressing variety of similar pages. For example, Google [1] lists the same pages in the top 5 for the query "Web link structure". Normally, users lose interest in checking variety of pages after checking first few pages. Providing variety pages of similar category enhances better use of this vast library. We can merge same pages into one and list only one.

Although hyperlinks (or citations) [5] and full-text are considered as the main key factor of the page ranking mechanism, degree of relevance and confidence of the pages can not be left out for a high quality page search. Confidence factor of the page plays significant role in page ranking. We

have defined the confidence value of the page as the probability of accessing the pages for the specific query the user is looking for from the past experience. It is a probabilistic value and represented by $C(\langle \text{feature} \rangle, \langle \text{Page} \rangle)$. We have applied the idea of Supermarket based association rule for the calculation of confidence. Making the assumption that most frequently accessed page by the users for specific query is the most relevant page and justified to assign high confidence value to the page for the query. In our method, we have calculated confidence value of all the pages against the most significant features of the page. For a popular Web site, million of users access the sites and million of queries are made in a day. Keeping track of such a large number of records poses a challenging problem. Most of the Web servers register such records in a Weblog. Weblog is considered as the best resource for this purpose.

Automated feature and category extraction method [5] gives high quality results. Users feedback, however, is also one major entity to determine the relevance of the page against the topic. Feedback based page relevance manipulation is reliable only if Web is accessed by genuine users. We have used automated features and categories extraction method and confidence of the pages for our study.

Category of the pages provides customize view of the pages and plays a significant role in relevant page searching. Some pages do not have a single occurrence of the topic, though they are highly related. For example, Yahoo! does not contain "Web Directory", but it is considered as one of the biggest Web directory. For the query "Web Directory", Yahoo! should also be placed into top list. Identifying most relevant feature and category poses a big challenge. Efficient automatic feature identification and categorization is proposed in [5]. In our method, not only citations and features, but also category is considered for page search.

Several studies find that multidimensional study is needed for finding relevant pages efficiently for specific topic. A broad study and centralize view can help in finding most relevant pages. A concept of virtual link can be used to get the centralized view. A virtual link can be defined as a links between pages having similar feature and/or characteristics. For example, Pepsi and Coca-Cola Company have the same characteristics, but it is unlikely that there will be citation between

them. But, a virtual link is assumed to exist between Pepsi and Coca-Cola. The practical application of virtual link graph includes identifying the category, focused search engines and centralized study of the Web nature. In our method, we minimize the full graph to the graph containing only relevant pages of our interest. All the pages not related to the search topic are pruned. Customizing the graph with only that of relevant pages can optimize the page ranking. We have considered citations, features, categories and confidence as the major criteria.

2. Our Approach

In our approach, the Web has been focused as a graph of pages having similar features and/or same category. The basic motivation of our work is that Web can be efficiently traversed in search of pages related with specific topic, if viewed from the search topic perspective. Before applying the ranking method, Web has been optimized for the specific topic, for which pages are being searched for. Our method basically involves four steps. First, we have extracted important features and categories for Web pages in the universe. For this purpose, we have downloaded significant number of pages from the Web through Google search engine. Second, prune the pages from the graph those are not relevant. Third, confidence of the pages for all the features and/or categories of each page are calculated. Fourth, using citations and confidence based recursive formula, page rank of all the pages for the specific query are calculated.

2.1 Features and Categories Extraction

There have been many efficient methods for extracting features and categories from pages. We have used full-text and extended anchor text method proposed by Glover et. al. [5] for extracting features and manipulated most relevant features from the past experience. For example, if a page P is accessed by the users for feature 'a' mostly than other features, then it is justified to assume feature 'a' as the most relevant feature of page P. In our case, the features, categories and relevance are equally taken into consideration and extracted both features and categories for all pages.

Relevance can be thought of as probability i.e., what is the probability that if the page P is accessed, then it will be accessed for the topic 'a'. Often there are many features in a page. The same

page can be accessed for other topic. It is also worthwhile determining the probability of accessing page P for the topic 'a' out of other features. We define extractor E as:

$E(P,a) = [N^2(P,a)/O(a)]/\sum[N^2(P,i)/O(i)]$, $\forall i$ in the set of features where $N(P,i)$ is total number of times page P is accessed for query 'i' and $O(i)$ is the total number of queries made for 'i'. It is also reasonable to claim that page having high $E(P,a)$ will be accessed next for the topic 'a'. $E(P,a)$ can be used to enhance most significant features and categories from the past experience. Most significant features and categories for each page are extracted from full-texts and extended anchor texts using the method proposed in [5] and past experience. We have ignored features and categories below a threshold value i.e., only the top features are selected.

In Figure 1, the label represents few most significant features and/or categories of the pages. The figure shows a sample graph ignoring the direction of the links, which is much different from reality. If we assume that the feature 'a' represents 'Steffi Graf', only the pages having feature 'Steffi Graf' can be selected as shown in Figure 2. If search topic is player, pages hit for player category can be considered. We don't distinguish between features and categories, since both can be considered for page searching.

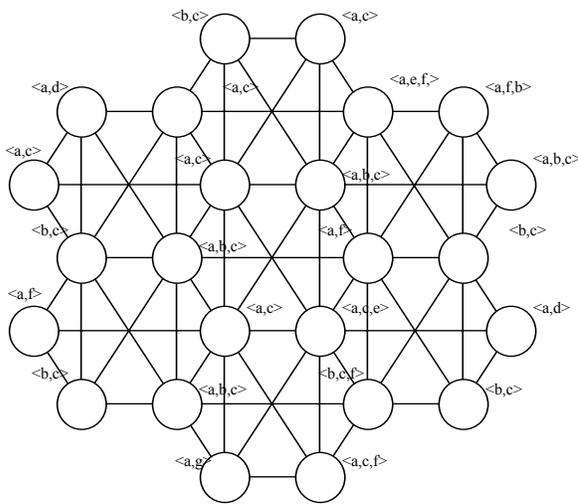


Figure 1: Graph containing features and categories.

2.2 Pruning Out the Irrelevant Pages

It is often found that pages, which do not have similar features, are also connected by hyperlinks.

For the topic Steffi Graf, we have found many pages, not related with Steffi Graf, but cited to the related pages. For example, in a personal home page, often there can be citation like 'my favorite player is Steffi Graf', citing to Steffi Graf's home page. In our method, such pages are used for extracting features and categories, but not taken into consideration for page rank calculation if the degree of corresponding feature is below the threshold value, considering as irrelevant page. Assuming that the citations from irrelevant pages are for some other features or interest, we do not consider such citations and pages. Such citations are useful for extracting category like 'player or 'Steffi Graf', but citing pages are not considered as the important pages for the search topic 'Steffi Graf'. We have found the relevant pages from the Web using Google and downloaded. Extracting features from the downloaded pages and searching the pages for each significant features using Google, citing pages are downloaded manually. At least 5 to 10 citations are taken for each page. In most of the cases, maximum number of citations are found to be within the relevant pages taken from the Google.

In Figure 2, all the irrelevant pages are pruned, and the citations from such pages are also removed.

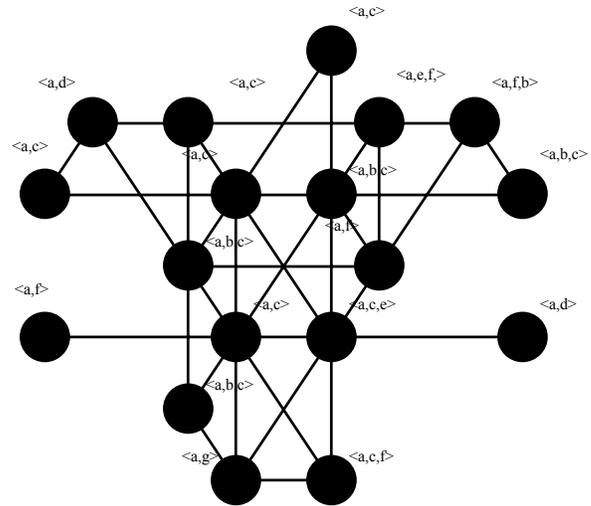


Figure 2: Pruned graph based on features and categories.

2.3 Confidence of a Page for a Specific Topic

Confidence of a page for a specific search topic is considered as one of the important measures in

our method. It is justified to minimize the search region to the customized graph, where customized graph is the graph containing similar features and/or categories. Confidence of a page, $C(a,P)$ is defined as the probability of accessing page P for the topic 'a.'

$C(a,P) = E(P,a) / \sum E(P_i,a), \forall P_i$ in the customized graph. Calculating $C(a,P)$ for the entire history is not realistic. Most recent records for the Web in demand will be better option for calculating $C(a,P)$. We have conducted our survey over a fixed period of time. For a popular Web site, millions of users access the sites and millions of queries are made in a day. Keeping tracks of such a large number of records is a challenging problem still exists today. Most of the Web servers register such records in a Weblog. It is considered as the best resource for this purpose.

2.4 Page Ranking using Citations and Confidence

We have customized the PageRank [2] only for the relevant pages. $PR(P)$ defined in [2] is redefined as $PR(A,a) = (1-d) + d [PR(T_1,a) / O(T_1) + \dots + PR(T_n,a) / O(T_n)]$, where T_i is the citing pages and $O(T_i)$ is the number of outgoing links from T_i for $i=1,2,\dots,n$ and 'd' is the damping factor i.e., what is the probability that a random user will get bored browsing page P .

The following relation of resultant page rank, $RPR(P,a) = PR(P,a) * C(a,P)$ gives significantly good result, if both, the PageRank and confidence are high. But, it reduces drastically in the case of new pages, which are cited by many relevant and high rank pages. Basic intuition is that such pages should have high page rank. Citation based search engines assign high page rank to such pages. By fixing a time period for calculating $C(a,P)$, above problem can be suppressed i.e., how many times page P has been accessed for the topic 'a' and how many times the query 'a' has been made in last one month. It also gives the demand of each pages in recent time. $C(a,P)$ is the probability of accessing page P for the topic 'a'.

We propose that, the damping factor 'd' defined by Brin & Page [2] can be substituted by $(1-C(a,P))$. Therefore, $PR(P,a) = C(a,P) + (1-C(a,P)) [PR(T_1,a) / O(T_1) + \dots + PR(T_n,a) / O(T_n)]$, where T_i is the incoming links to page P and $O(T_i)$ is the number of outgoing links from T_i for $i = 1,2,\dots,n$. $PR(P,a)$ can be calculated recursively

and $C(a,P)$ can be calculated in run time. By customizing the graph in question only for the specific feature enhances centralized view offering quality result.

3. Experimental Procedures and Results of our Methodology

For the experimental verification, we have selected most frequently requested queries from a survey conducted during the period of 5th to 14th January 2003 in our departmental Advanced Computing Lab. Top seven queries have been chosen. A sample set of top 30 pages for each top query is downloaded from Google. A base graph of the top 30 pages for each query is constructed. The links between each node in the base graph are made. We have manually found out 3 to 10 incoming citations to and 3 to 10 outgoing citations from the base nodes using random search in Google and expanded the graph up to the depth of 2 downloading the pages from Google. This way, we have constructed the graph of relevant pages for each topic. The expanded graph contains 500 to 800 nodes. Table 1 shows the seven top queries and corresponding number of requests submitted by the users during the survey period.

| Query | Number of requests submitted |
|---|------------------------------|
| "Database Management System" | 52 |
| "Java Programming" | 49 |
| "Linux Programming" | 37 |
| "Mobile Computing" | 36 |
| "Internet Service Provider" | 21 |
| "Relational Database Management System" | 19 |
| "Soft Computing" | 17 |

Table 1: Top 7 query topics and corresponding numbers of requests by users during the survey.

From the same survey, we had recorded all the pages that users had accessed for each query. In most of the cases, it is found that users access very few numbers of the pages in the lower label. Table2 shows the percentage of accessing the pages in each label by users during the survey conducted.

From the survey, it is observed that higher ranked pages are not always accessed more number of times. For example, in Table6 the page www.newriders.com/books/opl/ebooks/0735710430.html has been accessed more no. of times than

other top ranked pages in Google. The top ranked page www.linuxplanet.com/linuxplanet/ has not been accessed for even once. Confidence of a page 'P' for the topic 'a', C(a,P) is calculated for all the pages in the graph.

| Query | Label 1 | Label 2 | Label 3 |
|---|---------|---------|---------|
| "Database Management System" | 70.8% | 9.6% | 0.62% |
| "Java Programming" | 75.2% | 6.1% | 0.5% |
| "Linux Programming" | 81.4% | 3% | 0% |
| "Mobile Computing" | 63% | 1% | 0% |
| "Internet Service Provider" | 70% | 2% | 0.5% |
| "Relational Database Management System" | 53.5% | 0% | 1% |
| "Soft Computing" | 61.4% | 5.3% | 0% |

Table 2: Percentage of accessing pages in each label by users during the survey.

| Page (P): http://java.sun.com/ | | | |
|---|------|--------|---------|
| Features ('a') | O(a) | N(P,a) | E(P,a) |
| "Java Programming" | 49 | 30 | 0.35100 |
| "Java Applet Programming" | 15 | 12 | 0.18345 |
| "JSP" | 10 | 8 | 0.12230 |
| "Java" | 11 | 8 | 0.11118 |
| "J2ME" | 7 | 6 | 0.09828 |
| "J2EE" | 5 | 5 | 0.09555 |
| "J2SE" | 2 | 2 | 0.03822 |

| Page (P): http://www.oracle.com/ | | | |
|---|------|--------|---------|
| Features ('a') | O(a) | N(P,a) | E(P,a) |
| "Database Management System" | 52 | 31 | 0.42298 |
| "Oracle" | 19 | 17 | 0.34813 |
| "Relational Database Management System" | 10 | 10 | 0.22887 |

Table 3: Most relevant feature and/or category of the pages from past experience (above survey) using E(P,a).

From the same survey, we have observed that pages have also been accessed for different queries. For example, www.oracle.com has been accessed for queries "Database Management System", "Oracle", "Relational Database Management System". Using the past history user oriented features and categories can be extracted. We have proposed E(P,a) to manipulate the most relevant features and / or categories shown in Table3.

Tables 4, 5 and 6 show top 10 results obtained from our proposed method and the results given by Google for specific query topic. The result obtained from the Tables 4, 5 and 6 will be changed

| Google | PR(P,a) |
|---|---|
| http://java.sun.com | http://developer.java.sun.com/developer/onlineTraining/Programming/BasicJava1/ |
| http://java.sun.com/docs/html/CodeConvTOC.doc.html | http://developer.java.sun.com/developer/onlineTraining/new2java/ |
| www.apl.jhu.edu/~hall/java/ | http://java.sun.com/docs/html/CodeConvTOC.doc.html |
| http://developer.java.sun.com/developer/onlineTraining/Programming/BasicJava1/ | http://geaosoft.no/javastyle.html |
| http://developer.java.sun.com/developer/onlineTraining/new2java/ | http://math.hws.edu/javanotes/ |
| http://math.hws.edu/javanotes/ | www.apl.jhu.edu/~hall/java/tes/ |
| www.phrantic.com/scoop/onjava.html | developer.apple.com/java/javatutorial/ |
| www.webdeveloper.com/java/java_programming_groends_up.html | www.javaranch.com/style.jsp |
| http://geaosoft.no/javastyle.html | www.javalinex.net/ |
| www.ibiblio.org/javafaq/courese/ | http://javaboutique.internet.com/ |

Table 4: Top 10 pages ranked by Google and our methodology for the query "java programming".

time to time based on the users demand. As a result, it shows the most demanded pages in recent time and rank them in top. C(a,P) represents the probability of accessing page 'P' for the query topic 'a'. We also observed from the experiment that pages with higher probability value had often ranked high.

| Google | PR(P,a) |
|--|---|
| www.mysql.com/ | www.cql.com/ |
| www.oracle.com/ | www.mysql.com/ |
| www.postgresql.org/ | www.oracle.com/ |
| www.webopedia.com/TERM/D/database_management_system_DBMS.html | www.acm.org/sigmod/databaseSoftware/ |
| www.cql.com/ | www.ispras.ru/~knizhnik/gigabase.html |
| www.maverick-dbms.org/ | www.cs.wisc.edu/~dbbook/ |
| www.empress.com/ | www.ispras.ru/~knizhnik/goods.html |
| www.macnuclide.com/ | www.logger.fsec.ucf.edu/meet/Edbms.htm |
| www.advantagedatabase.com/ | http://s2kftp.cs.berkeley.edu:8000/mariposa/ |
| www.ispras.ru/~knizhnik/goods.html | www.iath.virginia.edu/elab/hfi0018.html |

Table 5: Top 10 pages ranked by Google and our methodology for the query "Database Management system".

| Google | PR(P,a) |
|--|--|
| www.linuxplanet.com/linuxplanet/ | www.advancedlinuxprogramming.com/downloads.html |
| http://leapster.org/linoleum/ | http://leapster.org/linoleum/ |
| www.advancedlinuxprogramming.com/ | www.newriders.com/books/opl/ebooks/0735710430.html |
| www.advancedlinuxprogramming.com/downloads.html | www.freelabs.com/~whitis/unleashed/ |
| www.ee.mu.oz.au/linux/programming/ | www.luv.asn.au/overheads/prog |
| www.pragana.net/ | http://sitereview.org/?article=777 |
| www.dwheeler.com/secure-programs/ | www.wrox.com/books/1861003013.htm |
| www.linuxlinks.com/portal/phpBB2/viewforum.php?f=8 | www.linuxlinks.com/portal/phpBB2/viewforum.php?f=8 |
| http://librenix.com/coding.php3 | www.pragana.net/ |
| www.luv.asn.au/overheads/prog/ | http://members.tripod.com/rpragana/ |

Table 6: Top 10 pages ranked by Google and our methodology for the query "Linux Programming".

4. Conclusion & Further Work

In this paper, we have described an approach for finding the probability of accessing a page for a specific query from the past experience. This probability can be used to enhance page ranking of the most significant pages in question. The graph in the Universe can be customized to the graph of related features and hence enhances extraction of most significant features and/or categories considering past experience, full-text and extended anchor text. $E(P,a)$ can be used for identifying most relevant categories. Customizing the graph to the graph of pages with only relevant features removes the citations from the irrelevant pages and hence uses only the relevant weight for calculating page rank. Using $C(a,P)$ and $PR(A)$ proposed by Brin and Page [2], we have evaluated the rank for the pages based on our small survey. It can be extended to larger area, say all over the world during certain period of time so as to get the most relevant and accurate result. $C(a,P)$ can easily be fitted to different page ranking algorithms proposed in [2, 8, 13, 14, 15, 16].

Apart from the full-text and anchor-text based most significant features identification mechanisms [5], the feedback collected from the genuine users can be one of the most significant measures. But, it is highly possible that users intentionally give high relevance feedback or random feedback without

proper judgments. Since we cannot assure genuine users, a sophisticated mechanism to identify genuine users is needed.

References

- [1] Google. Google search engine. <http://www.google.com>
- [2] Sergey Brin, Lawrence Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine," In *Proc. 7th Int. World Wide Web Conf.*, 1998.
- [3] Gary William Flake, Steve Lawrence, C. Lee Giles, Frans M. Coetzee. "Self-Organization and Identification of Web Communities," *IEEE Computer*, 35(3), 2002, pp 66-71.
- [4] Jon Kleinberg, Steve Lawrence. "The Structure of the Web," *Science Vol. 294*, 30 November 2001.
- [5] Eric J. Glover, Kostas Tsioutsoulouklis, Steve Lawrence, David M. Pennock, Gary W. Flake. "Using Web Structure for Classifying and Describing Web Pages," *WWW2002, Honolulu, Hawaii, USA, 7-11 May 2002*.
- [6] Steve Lawrence. "Context in Web Search," *IEEE Data Engineering Bulletin*, Vol. 23, November 2000, pp 25-32.
- [7] Réka Albert, Hawoong Jeong, Albert-László Barabási. "Diameter of the Web," *Nature*, Vol. 401, 9 September 1999.
- [8] Soumen Chakrabarti, Byron E. Dom, Ravi Kumar, Prabhakar Raghavan, Shidhar Rajagopalan, Andrew Tomkins, David Gibson, Jon Kleinberg. "Mining the Web's Link Structure," *IEEE Computer*, (32)8: August 1999, pp 60-67.
- [9] Steve Lawrence, C. Lee Giles. "Searching the World Wide Web," *Science*, Vol. 280, 3 April 1998.
- [10] Raymond Kolasa, Hendrik Blockeel. "Web Mining Research: A Survey," *ACM SIGKDD*, July 2000.
- [11] Dimitris Achlioptas, Amos Fiat, Anna R. Karlin, Frank McSherry. "Web Search Via Hub Synthesis," *IEEE Symposium on Foundations of Computer Science*, 2001.
- [12] Michael L. Mauldin. "Lycos: Design Choices in an Internet Search Service," *IEEE Expert*, January-February 1997, pp 8-11.
- [13] R. Lempel and S. Moran. "The stochastic approach for link-structure analysis (SALSA) and the TKC effect," In *Proc. 9th International World Wide Web Conf.*, 2000.
- [14] Alta Vista. Altavista search engine. <http://www.altavista.digital.com>.
- [15] Excite. Excite search engine. <http://www.excite.com>
- [16] Northern Light. Northern Light search engine. <http://nlsearch.com>.