Category-based PageRank Algorithm

B.Jaganathan[a], Kalyani Desikan[b]

[a,b]*Division of Mathematics,School of Advanced Sciences,V.I.T University,Chennai 600127,India*

Abstract

The dramatic growth of web information technology is forcing modern web search engines to look beyond the content of pages in providing relevant answers to queries. This paper has a brief description of the original PageRank algorithm, used in the Google search engine. This algorithm greatly improves the results of Web search by utilizing the link structure of the web. We propose an improved method for the computation of page rank on the basis of Categories of web pages. We experimentally compare the page ranks obtained using the original page rank method and our proposed category-based page rank method.

Keywords: World Wide Web; Search Engines; Information Retrieval; PageRank.

1. Introduction

With the rapid development of world-wide web, Internet has become the world's richest and most dense source of information. The users face the problem of getting the most relevant and useful information from the large collection of disordered information. Search engines are an important tool for users to retrieve information for a particular query. However, current search engines cannot fully satisfy the user's need for high-quality information search services; and it raises many new challenges for information retrieval.

Currently, the most classic Web structure algorithm is PageRank algorithm [1] that Sergey Brin and Larry Page proposed at Stanford University. In order to verify the performance of the algorithm, they successfully applied it to the Google search engine prototype, and now Google has become the world's most well-known search engine. However, the PageRank algorithm makes use of the hyperlinks-based structural analysis for measuring the relative importance of web pages, completely ignoring factors like content, topic and relevancy. It is difficult to achieve most relevant page /more informative page from the large collection of web pages. This paper discusses the PageRank algorithm and its computation formula. We propose a new technique for the computation of page rank of web pages segregated into categories based on user interests.

1.1. Web Page Ranking Algorithm

The size of the www is growing rapidly and at the same time the number of users has also grown incredibly. With increasing number of users on the web, the number of queries submitted to search engines is also growing exponentially. Therefore the search engines must be able to process these queries efficiently. Web ranking techniques are applied in order to extract only relevant documents from the database and provide the intended information to the users.

The ranking algorithms are based on different nature of searching and some of them are

- Link analysis algorithm
- Personalized web search ranking algorithm
- Page segmentation algorithms

In the following sections, we briefly discuss the above mentioned algorithms.

1.2. Link Analysis Algorithm

The link analysis algorithms are based on the link structure of the web pages. The quality of results from search engines is generally lower than what the user expects and this quality can be improved greatly if pages are ranked according to some criteria based on links between the pages.

ISBN 978-93-82338-91-8 © 2013 Bonfring

1.3. Personalized Web Search Ranking Algorithm

For a given query, a personalized Web search can provide different search results for different users or organize search results differently for each user, based upon their interests, preferences, and information needs. Personalized web search differs from generic web search, which returns identical research results to all users for identical queries, regardless of varied user interests and information needs.

1.4. Page Segmentation Algorithms

There are several applications of webpage segmentation. Segments demarcate informative and non-informative content on a web page. They can also discriminate between different types of information. This is very useful in web ranking. Consider a multiword query whose terms match across different segments in a page, this information can clearly be useful in adjusting the relevance of the page to the query.

We now discuss the PageRank algorithm originally proposed by Larry Page and Sergey Brin.

2. Page rank algorithm

PageRank method for computing a ranking for every Web page based on the graph structure of the web has applications in search, browsing and so on.

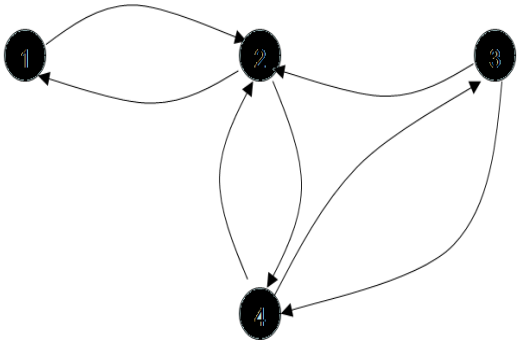

Fig: 1 Web graph with hyperlink

The web graph depicts the directed links between pages of the World Wide Web. A graph, in general, consists of several vertices, some pairs connected by edges. In a directed graph, edges are directed lines or arcs between vertices. The web graph is a directed graph, whose vertices correspond to the pages of the WWW, and a directed edge connects page X to page Y if there exists a hyperlink on page X, referring to page Y. Also a hyperlink can be classified as an in-link or out-link. For web page W, an in-link is a hyperlink from another web page which contains a link pointing to W. To web page W, an out-link is a hyperlink appearing in W which points to another web page. The essential idea behind the PageRank algorithm can be described as follows. A page has a high rank if the sum of the ranks of its in-links is high. This covers both the cases when a page has many in-links and when a page has a few highly ranked in-links. The PageRank computation formula can be formally defined as follows. Let u be a web page. Let be the set of pages that point to u, O (u) be the number of out-links from u. According to the PageRank algorithm [1], an iteration of the page rank value PR (u) of the page u can be computed using the following formula:

$$PR(u) = (1-d) + d \sum_{v \in B_u} \frac{PR(v)}{O(v)} \quad (1)$$

ISBN 978-93-82338-91-8 © 2013 Bonfring

The PageRank algorithm works on the principle that an imaginary surfer who randomly clicks on links will eventually stop clicking. The probability, at any step, that the person will continue to click is known as the damping factor, denoted as d. In the above formula, the parameter d is the damping factor whose value is between 0 and 1, and it is usually set to 0.85.

3. Category-based page rank

3.1. Category-based PageRank Method

Generally users search for web pages within a particular category/topic only. For example individuals normally search the web for information related to say a particular topic or category of information like music, movies, and different kinds of sports etc. It depends on an individual's interest. Hence, it would be ideal to calculate the page rank based on categories of web pages. Before calculating the page rank of each page, the pages must be first grouped into different categories/topic using web clustering techniques. It is reasonable to assume that the probability that a user moves from one web page to another within his category of interest is high compared to the probability of him navigating across topics/categories.

In this paper, assuming that the web pages are clustered based on categories/topics, we propose an algorithm to calculate the page rank taking into account the category of the web page. In our algorithm, we have considered a higher damping factor for navigating between web pages within a category compared to navigating across categories. We have also introduced the following terminology:

- A link within a category is called an **intra-link**. For an intra-link the value of the teleport probability (damping factor) is taken as 0.85.
- A link between different categories is called an **inter-link**. For an inter-link the value of the teleport probability is 0.15.

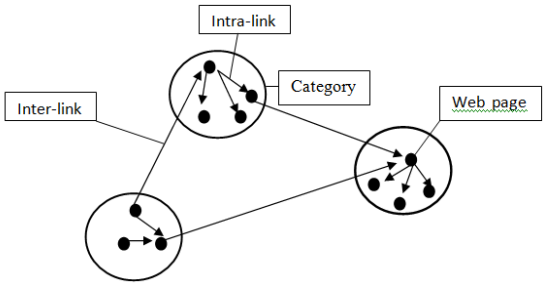

Fig: 2 Category-based web graph

Given below is our proposed category-based PageRank algorithm

- BEGIN
- Initialization (concurrently on the web pages for each Category)
  - Initialize all page weights to 1
- Calculation of page rank
- Compute the page rank value *PR (u)* of the page *u* using the formula

$$PR(u) = (1-d) + d \sum_{v \in B_u} \frac{PR(v)}{O(v)} + d^* \sum_{w \in B_u^*} \frac{PR(w)}{O(w)} \quad (2)$$

where B_u is the set of web pages within the same category as *u* that point to *u* and B_u^* is the set of web pages that point to *u* from another category. Let *d* and d^* be the damping factor of intra link and inter link respectively.

ISBN 978-93-82338-91-8 © 2013 Bonfring

- Repeat the above step until the page rank values converge.
- END

4. Comparison between the page rank algorithms

Given below are two sample graphs showing web pages grouped based on three different categories.

a b

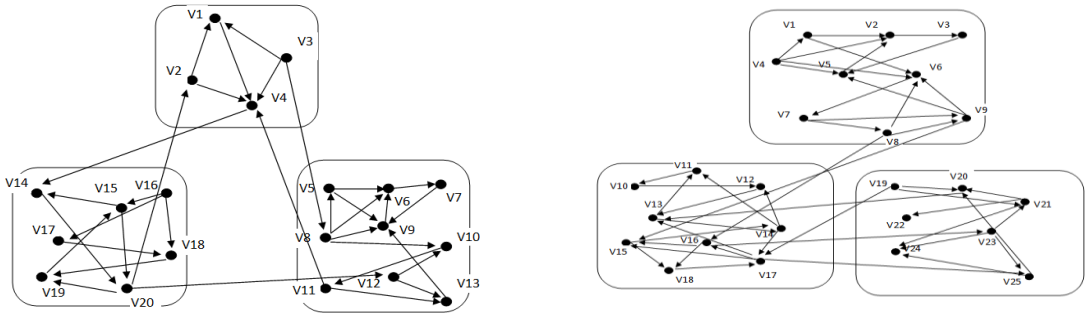

Fig: 3. (a) Web graph-1; (b) Web graph-2

We calculated the page rank for the web pages for the two web graphs using the original Page rank algorithm and our proposed algorithm. The tables given below shows the page ranks computed for the web pages using the two algorithms. The web pages (vertices) are arranged in the tables in the increasing order of page rank value.

Table 1. Comparison between the page rank algorithms for web graph -1.

Web Page	Category-based PageRank	Web Page	PageRank
v9	2.0365	v9	2.4446
v6	1.9934	v6	2.3504
v7	1.8444	v7	2.1478
v19	0.7432	v20	1.8103
v20	0.6836	v14	1.5974
v11	0.6199	v4	1.3469
v10	0.5528	v19	1.2221
v4	0.5454	v11	0.9458
v15	0.5086	v12	0.9373
v13	0.4877	v10	0.9362
v12	0.3814	v2	0.9194
v14	0.3736	v13	0.8163
v18	0.3561	v15	0.7119
v1	0.278	v1	0.5832
v2	0.2013	v18	0.3561
v17	0.1925	v8	0.1925
v5	0.1835	v17	0.1925
v8	0.1575	v5	0.1909
v3	0.15	v3	0.15
v16	0.15	v16	0.15

Table 2. Comparison between the page rank algorithms for web graph -2.

Web Page	Category based PageRank	Web Page	PageRank
v5	2.3059	v5	2.5431
v2	2.237	v2	2.4379
v3	2.0514	v3	2.2222
v13	1.1953	v13	1.9569
v12	1.1624	v12	1.6422
v9	1.015	v11	1.3619
v6	0.9736	v14	1.3405
v7	0.9256	v9	1.3395
v10	0.9118	v10	1.3086
v11	0.8956	v7	1.2628
v14	0.8373	v6	1.2471
v17	0.7336	v17	1.0952
v8	0.5549	v15	0.8437
v15	0.4217	v8	0.7681
v18	0.3942	v18	0.5899
v20	0.3822	v24	0.4852
v24	0.3642	v20	0.4834
v16	0.3059	v25	0.4483
v21	0.2595	v16	0.3828
v22	0.2235	v21	0.2793
v25	0.2233	v23	0.2313
v1	0.2138	v22	0.2291
v23	0.1615	v1	0.2138
v4	0.15	v4	0.15
v19	0.15	v19	0.15

ISBN 978-93-82338-91-8 © 2013 Bonfring

The graphs below shows the page rank values obtained using the two algorithms for Web graph 1&2.

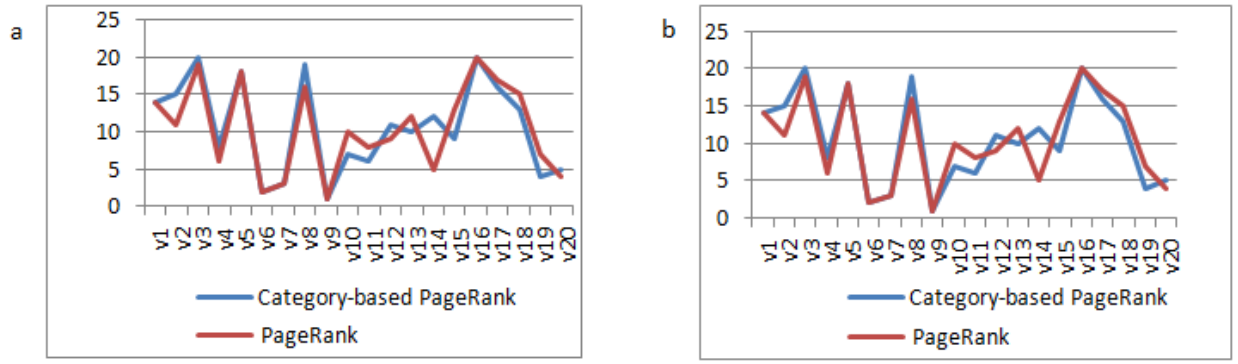

Fig: 4. (a) Page rank value for the web graph -1 (b) Page rank value for the web graph -2

It can be seen that the page ranks computed using the two algorithms agree with respect to the top three web pages for both the sample web graphs.

5. Conclusion/results and discussion

It is reasonable to assume that a user who searches for particular information will navigate between pages belonging to the category/topic of his interest. Hence, in this paper, we have made use of this fact in calculating the page ranks based on user's interests.

References

[1] Page, L., Brin, S., Motwani, R., Winograd. T.,, 1998. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project.
[2] Kleinberg, J.M., 1999. "Authoritative sources in a hyperlinked environment," Journal of the ACM, 46 (5), pp. 604–632.
[3] Dilip Kumar Sharma, Sharma A. K., 2010 ."A Comparative Analysis of Web Page Ranking Algorithms" Journal on Computer Science and Engineering, Vol. 02, No. 08, pp.2670-2676.
[4] Laxmi Choudhary., Bhawani Shankar Burdak., 2012. "Role of Ranking Algorithms for Information Retrieval" International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.4,pp.21-34.
[5] Mandar Kale ., Santhi Thilagam, P., 2008."DYNA-RANK: Efficient calculation and updation of PageRank," International Conference on Computer Science and Information Technology,
[6] Sepandar, D., Kamvar Taher, H., Haveliwala Christopher, D., Manning Gene, Golub H., 2003. "Exploiting the Block Structure of theWeb for Computing PageRank," Stanford University Technical Report.

ISBN 978-93-82338-91-8 © 2013 Bonfring