ISSN: 2277-3754
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 1, January 2012

Salutary Web Service Retrieval

Nistha Goyal, Suneet Kumar, Arjun Arora, Bipin Chandra
nistha_goel@yahoo.com,suneetcit81@gmail.com,arjunarora06dit@gmail.com,
bipin.updhyay88@gmail.com

Abstract*— Today Web Users generally rely on service retrieval systems or search engines in order to obtain any form of service from the web. Thus it becomes essential for them to provide good quality results to users which contain salutary information that best fulfills user's requirements and contain data that is trustworthy and contain least insignificant information. As the web is open for everyone without any restrictions so there are possibilities that the data retrieved contains false facts deceits and assumptions and thus the service retrieval systems are becoming more aware about the authenticity and significance of the web documents to provide best service retrieval. Web Mining is widely used in this context which is used to discover the content of the web, the user's behavior in the past and the web pages that the users want to view. Google; one of the most popular search engine uses the concept of Page Rank in order to rank web documents so to maintain the trustworthiness and importance of the documents and to retrieve results. Page Rank is used in Web Structure mining; a sub category of Web Mining which analyze link structure of the hyperlinks between the documents and then rank the pages on the basis of that structure. In this paper an Optimal Service Retrieval Algorithm is proposed that uses the concept of Page Rank and deals with the ranking of documents and optimization of the ranking procedure. The paper has also analyzed the behavior of proposed algorithm for different values of Clamminess Element.

***Index Terms*— Web Mining, Link-Structure, Random Surfer Model, Markov Chain, Service Retrieval System.**

I. INTRODUCTION

With the wide use of the web in various fields like e-commerce, e-learning, e-news, finding user's need and providing useful information or services are the major factors of consideration. The purpose of service retrieval is to facilitate user's access to service or information that is relevant to his service needs. A salutary service retrieval system should provide the user with easy access to the useful service in which he is interested. Basically a service retrieval system matches user queries to documents stored in database. The data stored in these databases is of unstructured nature which means that the data does not have clear, semantically obvious structure [4], thus a service retrieval system has to search recover and understand data from large collection of stored unstructured data. A service retrieval system can adopt any of the various available methods like keyword based search, concept based search, refined search in order to retrieve service. The web contains huge amount of truth and lies, assumptions and contradictions thus it becomes the for most requirement of a service retrieval system to use techniques that provides a criteria to identify data that best matches user's needs and also assures trustworthiness of the retrieved data. Web Mining helps in fulfilling these requirements. Web mining is used to discover the content of the Web, the users' behavior in the past, and the WebPages that the users want to view in the future. Web mining consists of Web Content Mining (WCM), Web Structure Mining (WSM), and Web Usage Mining (WUM) [10]. WCM deals with the discovery of useful information from web content. WSM discovers relationships between web pages by analyzing web structures. WUM ascertains user profiles and the users' behavior recorded inside the web log file Based on the topology of the hyperlinks, WSM categorizes web pages and generates related patterns, such as the similarity and the relationships between different Web site. This paper focuses on WSM , Page Rank algorithm which is commonly used in Web Structure Mining and on the ranking problem according to which every owner of the document wants to improve the ranking of its document for that it can do many manipulations in its document like increasing the number of links to the document or page by the dummy pages[3].Thus ranking of a page plays a significant role in the process of service retrieval and this paper introduces an algorithm for ranking the web pages by using web graph and also analysis its behavior.

The rest of this paper is organized as follows. Section 2 presents a brief background review of the Ranking Process as followed by Google. Section 3 describes the various challenges faced by the current. Section 4 presents an algorithm that computes the relative ranks and shows the relative level of trust between web pages.

II. RELATED WORK

With the rapid growth of the Web, providing relevant pages of the highest quality to the users based on their queries becomes increasingly difficult. The reasons are that some web pages are not self-descriptive and that some links exist purely for navigational purposes. Therefore, finding appropriate pages through a search engine that relies on web contents or makes use of hyperlink information is very difficult. To address the problems mentioned above, several algorithms have been proposed. Among them are Page Rank [17] and Hypertext Induced Topic Selection (HITS) [10] algorithms. Page Rank is a commonly used algorithm in Web Structure Mining. It measures the importance of the pages by analyzing the links [3]. Page Rank has been developed by Google and is named after Larry Page, Google's co-founder and president [17]. Page Rank ranks pages based on the web structure. Google first retrieves a list of relevant pages to a given query

ISSN: 2277-3754
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 1, January 2012

based on factors such as title tags and keywords. Then it uses Page Rank to adjust the results so that more "important" pages are provided at the top of the page list [17]. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The seminal papers [3] introduced Link Analysis Ranking, where hyperlink structures are used to determine the relative authority of a Web page and produce improved algorithms for the ranking of Web search results. In the current era there is much concern in using random graph models for the Web. [1] Describes that there are two models: One is the Random Surfer model, introduced by [5], and the Page Rank-based selection model, proposed by [12]. Page Rank-based selection model tries to capture the effect that these search engines have on the growth of the Web by adding new links according to Page rank. The Page Rank algorithm is used in the Google search engine [14] for ranking search results. [2] Illustrate Google is designed to be a scalable search engine.

According to [4, 6] Page Rank is defined as the stationary state of a Markov chain. The chain is obtained by perturbing the transition matrix induced by a web graph with a damping factor that spreads uniformly part of the rank. The choice of damping factor (α) is eminently empirical, and in most cases the original suggestion $\alpha = 0.85$ by [3] is still used. Recently, however, the behavior of Page Rank with respect to changes in α was discovered to be useful in link-spam detection [9]. Moreover, an analytical justification of the value chosen for α is still missing. [6] give the first mathematical analysis of Page Rank when α changes. In particular, [6] shows that, contrarily to popular belief, for real-world graphs values of α close to 1 do not give a more meaningful ranking. [11] Determines the order in which to display web pages, the search engine Google computes the Page Rank vector, whose entries are the Page Ranks of the web pages. The Page Rank vector is the stationary distribution of a stochastic matrix, the Google matrix. The Google matrix in turn is a convex combination of two stochastic matrices: one matrix represents the link structure of the web graph and a second, rank-one matrix, mimics the random behavior of web surfers and can also be used to fight web spamming. As a consequence, Page Rank depend mainly the link structure of the web graph, but not on the contents of the web pages. [1] says, the Page Rank of the first vertex, the root of the graph, follow the power law. However, the power undergoes a phase-transition as parameters of the model vary.

III. CHALLENGES IN WEB SERVICE ANALYSIS AND RETRIEVAL

A critical goal of successful information retrieval on the web is to identify which pages are of high quality and relevance to a user's query. There are many aspects of Web Service Retrieval that differentiate it and make it somewhat more challenging than traditional service retrieval systems. One particularly intriguing problem in web IR arises from the attempt by some commercial interests to unduly heighten the ranking of their web pages by engaging in various forms of spamming [19]. One common method of spamming involves placing additional keywords (or even entire dictionaries) in invisible text on a web page so that the page potentially matches many more user queries, even if the page is really irrelevant to these queries. Decentralized content publishing is the main reason for the explosive growth of the web. Corresponding to a user query there are many document that can be retrieve by search engine. And every owner of the document wants to improve the ranking of its document for that it can do many manipulations on its document like increasing number of the link to the page by the dummy pages. Commercial search engine have to maintain the integrity of their search results that's why effort made by them are not publicly available. Democratization of content creation on the web generates a new challenge in Web Service Retrieval System. This meant that the web contained truth, lies, contradictions and suppositions on a grand scale. This gives rise to the question: which web pages do one trust? In a simplistic approach, one might argue that some publishers are trustworthy and others not begging the question of how a search engine is to assign such a measure of trust to each website or web page. One more challenge is that a Fast crawling technology is needed to gather the web documents and keep them up to date.

IV. OPTIMAL SERVICE RETRIEVAL ALGORITHM

Page Rank of a document can be defined as the fraction of time that the surfer spent on that document on the average. It can also be defined as a technique for Link Analysis that assigns a numerical score to each document stored on the web on the basis of which a ranked list of results is provided to the user for his queries. If a random surfer starts at a web page and executes a random walk on the web by proceeding from his current page to a randomly chosen page that his current page hyperlinks to and proceeds in this manner from node to node then he visits some nodes more often than others; intuitively these are the nodes with many links coming in from other frequently visited nodes the idea behind Page Rank is that pages visited more often in this walk are more important. Thus, the probability that the random surfer visits a document is its Page Rank [1]. And the probability that random surfer will get bored and restart from some another random document is the Clamminess Element (say C) and with the probability of (1-C) follow the out link chosen randomly [1]. [4] Describes that the Markov Chain is a discrete-time stochastic process: a process that occurs in a series of time-steps in each of which a random choice is made. There is one state corresponding to each web page. Hence, a Markov chain consists of N states if there are N no of web pages in the collection. A Markov chain is characterized by an $N \times N$ transition probability matrix P each of whose entries is in the interval [0, 1]; the entries in each row of *P* add up to 1. The Markov Chain can be in one of the N states at any given time step; then, the entry P*ij* tells us the probability that the state at the next time step is j, conditioned on the current state being i. Each entry *Pij* is known as a transition probability and

ISSN: 2277-3754
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 1, January 2012

depends only on the current state *i*; this is known as the Markov property.

A Markov chain's probability distribution over its states may be viewed as a probability vector: a vector all of whose entries are in the interval [0, 1], and the entries add up to 1. According to [4, 13] the problem of computing bounds on the conditional. Steady-state probability vector of a subset of states in finite, argotic discrete-time Markov chains is considered. Features of Page Rank Algorithm are:

- Page Rank Algorithm computes Page Rank values Offline by considering the Link structure of the whole web and using the Web Graph.
- It is Independent of user's queries and assigns a value to every document on the web.
- This algorithm rank the pages individually and not the website as a whole.
- It is concerned with static quality of a web page.
- Page Rank is a model of user's behavior.

A. Algorithm and Implementation

This algorithm basically considers each web document and the linking between them as a Web Graph where each node of the graph represents a web document and each edge of the graph represents an out link from one document to another. The algorithm takes this web graph as an input and then assigns a rank to every document which can specify the relative authorization of that document on the web. In the proposed algorithm, N is the Number of documents in the collection. "P" represents the probability Transition Matrix. C is the Clamminess Element and x is the probability vector. The Optimal Service Retrieval Algorithm which is based on Page rank Concept is given below.

1) Construct a Web Graph.
2) Compute number of out links from a particular node say Freq.
3) Calculate a NxN Matrix 'M[i] [j]'
//where N is the number of web pages or web documents in the web graph.

```
For i=1 to N
   For j=1 to N
    //if node i has no out link
       if (freq==0) then
        M[i][j]=0
      else
        M[i][j]=1/Freq
```

4) Multiply the resulting matrix by (1- C).
5) Add C/N to every entry of the resulting matrix to obtain the *Probability Transition Matrix*.
For all i,j€1 to Ndo

```
P[i][j]=(M[i][j]*(1-C))+(C/N);
```

6) Randomly select a node from 0 to N-1 to start a walk say s_int .
7) Initialize Random surfer and it to keep account of number of iterations required to 0.
8) try to reach at steady state within 200 iterations otherwise toggling occur
9) Multiplying probability transition matrix with probability vector to get steady state.
10) Check either system enter in steady state or not
11) Print the ranks stored in Probability vector X and exit.

B. Assumptions used

If there are multiple links between two pages, only a single edge is placed.
- No self loops are allowed.
- The edges could be weighted, but we assume that no weight is assigned to edges in the graph.
- Links within the same web site are removed.
- Isolated nodes are removed from the graph.

C. Random Surfer and State Description

This implementation is basically based upon [4] random surfer model and Markov chain. Actually the random surfer have to visit web graph's node according to some distribution on the bases of that random surfer can be at any time in one of the following four possible states. Initial state is that state of the system from where it will start its walk. We set the system in the random state by randomly selecting a node using random function and set corresponding to that node 1 in the probability vector. Rest of the values in the probability vector is zero. Steady state is that state of the system when the probability vector of random surfer fulfills the properties of irreducibility and aperiodicity's. To check either the system get the steady state or not two successive values of the probability vector must be same. Ideal state is that state of the random surfer when the system achieves the steady state but at the same time. Page Ranks are distributed uniformly to all documents. Toggling state is achieved by the random surfer when the system not able to reach at steady state and just toggle between two set of Page Ranks.

D. Results and Discussion

Graph structures used for analysis affect the performance of the algorithm. [18] Examine how different structures in the graphs affect their performance. The following web graph is used for analyzing various factors of explained algorithm.

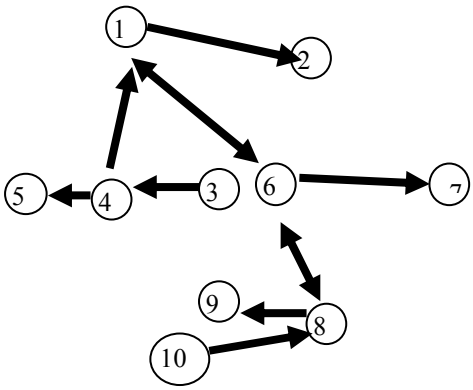

Fig1. Web Graph Used For Analysis

ISSN: 2277-3754
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 1, January 2012

To analyze the convergence speed we collect the information of number of iterations required to random surfer to reach at a steady state and the corresponding graph is shown in fig 2.In fig 2 infinity value is shown by 250.

Table 1: Moister Factor & No. of Iterations

Moister Factor	No. Of Iterations
0	Infinity
0.05	Infinity
0.1	159
0.15	105
0.2	Infinity
0.25	Infinity
0.3	61
0.35	Infinity
0.4	34
0.45	Infinity
0.5	33
0.55	24
0.6	23
0.65	22
0.7	18
0.75	18
0.8	18
0.85	12
0.9	12
0.95	10
1	2

Table 2: DOC ID & Rank Provided

DOC ID	1	2	3	4	5	6	7	8	9	10
Rank	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1

From the above table we analyze that Convergence speed decreases as the moister factor move from 0 to 1, therefore damping factor must be selected closer to 1 from the point of convergence speed. Second problem analyzed is as the Moister Factor is 1 Random Surfer enters into the ideal state and the corresponding graph is shown in fig 3. Here DOC ID is the Integer number assigned to every web document.

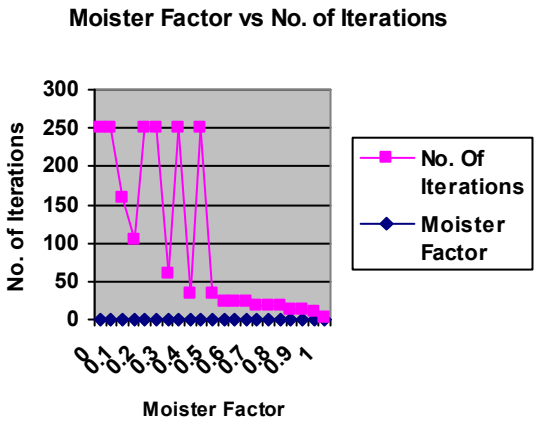

Fig. 2. Graph between Moister factor and No. of Iterations

The following graph shows that no rank is provided when the Moister Factor less than .5 and system enter into the toggling state most of the time from the above tables and analysis it is concluded that the Clamminess Element plays main role in this algorithm and performance of algorithm can be improved if this factor is selected properly. The value of Clamminess Element can vary from 0 to 1 but in most of the cases system enter into the toggling state if value selected is less than 0.5 and at the value 1system enter into ideal state giving insignificant results. Value must be closer to 1 but cannot be 1. As shown in fig 2 systems achieve a steady state in less no. of iterations if Clamminess Element value is closer to 1.

V. Conclusion

The current study was conducted to demonstrate how the link structure of the web can be used to provide the ranking to various documents. This ranking can be provided offline. With the help of this approach one can prioritize the various documents on the web independent of the query. However a complete score computation is based on various other factors. In the proposed algorithm a damping factor is used that play a very important role on the analysis of the algorithm. After the analysis it is concluded that damping factor must not be selected closer to zero. At the damping factor one, the system enters into the ideal state and the ranking provided is insignificant. As per evaluation the damping factor must be selected greater than or equals to 0.5. However, if we consider convergence speed as only factor to evaluate the performance than the best moister factor will be .95. The proposed algorithm is query independent algorithm and does not consider query during ranking.

Acknowledgment

I would like to express my gratitude to all the people who have given their heart welling support in making this completion a magnificent experience. And a special acknowledgement to the authors of various research papers and books which help me a lot.

ISSN: 2277-3754
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 1, January 2012

REFERENCES

[1] Prasad Chebolu, Páll Melsted," Page Rank and the random surfer model"; Symposium on Discrete Algorithms Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms; Pages: 1010-1018.Year : 2008.

[2] Sehgal, Umesh; Kaur, Kuljeet; Kumar, Pawan,"The Anatomy of a Large-Scale Hyper Textual Web Search Engine"; Computer and Electrical Engineering, 2009. ICCEE '09. Second International Conference on Volume 2, 28-30 Dec. 2009 Page(s):491 - 495; Year 2009.

[3] Sergey Brin, Lawrence Page, "The anatomy of a large-scale hyper textual Web search engine", Proceedings of the seventh international conference on World Wide Web 7, p.107-117, April 1998, Brisbane, Australia.

[4] Christopher D. Manning, Prabhakar Raghavan Hinrich Schütze;" An Introduction to Information Retrieval"; Publisher: Cambridge University Press New York, NY, USA, Pages: 461-470 Year of Publication: 2008.

[5] Blum, T.-H. H. Chan, and M. R. Rwebangira. ;"A random-surfer web-graph model". In ANALCO '06: Proceedings of the eight Workshop on Algorithm Engineering and Experiments and the third Workshop on Analytic Algorithmic and Combinatory, pages 238-- 246, Philadelphia, PA, USA, 2006. Society for Industrial and Applied Mathematics.

[6] Paolo Boldi, Massimo Santini, S. Vigna; " Page Rank as a Function of the Damping Factor" ; International World Wide Web Conference Proceedings of the 14th international conference on World Wide Web Chiba, Japan pages: 557 - 566 Year of Publication: 2005.

[7] Prasad Chebolu, Páll Melsted," Page Rank and the random surfer model"; Symposium on Discrete Algorithms Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms; Pages: 1010-1018.Year : 2008.

[8] R. Lempel, S. Moran;" Rank-Stability and Rank-Similarity of Link- Based Web Ranking Algorithms in Authority-Connected Graphs" Publisher**:** Kluwer Academic Publishers, April 2005 Information Retrieval, Volume 8 Issue 2, Pages: 245 - 264; Year: 2005.

[9] Hui Zhang, Ashish Goel, Ramesh Govindan, Kahn Mason, and Benjamin Van Roy. "Making eigenvector-based reputation systems robust to collusion". In Stefano Leonardi editor, Proceedings WAW 2004, number 3243 in LNCS, pages 92–104. Springer-Verlag, 2004.

[10] S. Pal, V. Talwar, and P. Mitra. Web mining in soft computing frame work: Relevance, state of the art and future direction IEEE Trans. Neural Networks, 13(5):1163–1177, 2002.

[11] Rebecca S. Wills;"Mathematical Properties and Analysis of Google's Page Rank "Ilse C.F. Ipsen, Year: 2006.

[12] Gopal Pandurangan, Prabhakar Raghavan, Eli Upfal, "Using Page Rank to Characterize Web Structure", Proceedings of the 8th Annual International Conference on Computing and Combinatory, page No..330-339, August 15-17, 2002.

[13] Turul Dayar, Nihal Pekergin, Sana Younès; "Conditional steadystate bounds for a subset of states in Markov chains" ACMInternational Conference Proceeding Series; Vol. 201 Proceeding from the 2006 workshop on Tools for solving structured Markov chains Article No.: 3 Year: 2006.

[14] ShuMing Shi, Jin Yu, GuangWen Yang, DingXing Wang;" Distributed Page Ranking in Structured P2P Networks"; Proceedings. 2003 International Conference on Publication Date: 9-9 Oct. 2003 On page(s): 179-186 Year: 2003.

[15] Zhanzi qui,Matthias Hemmje,Erich J.Neuhold,"Using Link types in web page ranking and filtering"; IEEE Computer Society Proceedings of the Second International Conference on Web Information Systems Engineering (WISE'01) Volume 1 **;** Page: 311 Year of Publication: 2001.

[16] C. Ridings and M.Shishigin.Pagerank uncovered. Technical report, 2002.

[17] Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, Panayiotis Tsaparas;"Link analysis ranking: algorithms, theory, and experiments" ACM Transactions on Internet Technology (TOIT) Volume 5, Issue 1 (February 2005) Pages: 231 - 297 Year: 2005.

AUTHOR BIOGRAPHY

Nistha Goyel Assistant professor in computer science deptt in Dev bhoomi Institute of technology, dehradun, India.

Suneet Kumar Assistant professor in computer science deptt in Dehradun Institute of technology, dehradun, India.

Arjun Arora Assistant professor in computer science deptt in Dehradun Institute of technology, dehradun, India.

Bipin Chandra Assistant professor in computer science deptt in Dehradun Institute of technology, dehradun, India.