International Journal of Science Engineering and Advance Technology

Implementing Efficient Search Results and Navigation in website

J.Prasanth kumar
M.Tech Scholar
Department of Computer Science and Engineering, Andhra Loyola Institute of Engineering and technology, Vijayawada(AP),India
Prasanth9001@gmail.com

ABSTRACT

Designing structured website and organizing search results in website is Challenging task. For this we propose mathematical programming model to improve user navigation as well as novel search interface that enables the user to navigate large number of query results by organizing them using the concept hierarchy. The results are organized into a navigation tree. At each node expansion step we implement classification algorithms [TF-IDF, SVM] for efficient classification of nodes and an advanced k-means algorithm for assigning priorities to the web site navigation links based on the efficient content retrieved. It reveals only a small subset of the concept nodes finally the expected user navigation cost is minimized.

Key Terms: Website Design, User Navigation, Page Ranking, Mathematical Programming Model.

Introduction:

Web graph properties are measured by considering the Web or a portion of it, such as a web site, as a directed hypertext graph where nodes represent pages and edges hyperlinks referred to as the Web graph. Web graph properties reflect the structural organization of the hypertext and hence determine the readability and ease of navigation. Poorly organized web sites often cause user disorientation leading to the “lost in cyberspace” problem. These metrics can aid web site authoring and create sites that are easier to traverse. Variations of this model may label the edges with weights denoting, for example, connection quality, or number of hyperlinks.

There are technique that discovers the gap between Web site designers' expectations and users' behavior. The former are assessed by measuring the inter-page conceptual relevance and the latter by measuring the inter-page access co-occurrence. The discovery of pages that are conceptually related but rarely co-occur in visits suggests areas where Web site design improvement would be appropriate. Further, the technique suggests how to apply quantitative data obtained through a multiple regression analysis that predicts hyperlink traversal frequency from page layout features.

The appearance and rapid development ofInternet has greatly changed environment ofinformation retrieval. However, the rankalgorithms for search engine based on Internetare directly related to experiences in usingwhen users perform information retrievals inthe new environment. The existing rankalgorithms for search engine are mainly basedon the link structure of web pages, and the twomain representative algorithms are PageRankalgorithm and HITS algorithm. Many scholarsand research institutions have made newexplorations and improvements based on thesetwo algorithms, and some mature integratedrank models suitable for search engines weregenerated.

In this paper, we study the shortcomings ofsearch engines, and provide further analysis onPageRank algorithm and Hits algorithm.Beside, we discuss the existing improvedalgorithms based on link structure, andprovide analysis on the improvement ideas ofexisting search engine rank technology.Moreover, research on traditional conceptsemantic similarity computation models basedon domain ontology is given as well.

According to the characteristics andshortcomings of existing models andalgorithms, we firstly propose an improvedconcept semantic similarity computationmodel. Then, an improved rank algorithmwhich integrating categorization technologyand traditional link analysis algorithm basedon it is given in this paper, which improvesHITS algorithm in two aspects, the preprocessingof Web pages and analysis on thelink

ISSN 2321-6905
January-2015

structure of Web page. At last, theevaluations are provided as well.

Search engines have gradually become a highefficient and convenient way for data queryand information acquisition to people. Withthe continuous development of search enginetechnology, the current mature commercialsearch engines have experienced severalgenerations of evolution. Meanwhile, Webinformation retrieval technology, which is theessence of search engines, includingcommercial products has come out for about20 years. In this period of time, greatprogresses in the aspects of retrieval keytechnology, system structure design, queryalgorithm and etc. are made, and a lot ofcommercial search engine services are beingused on Web. Compare with these progresses,the rapid increment of data on Web weakensthe achievement obtained in the research fieldof Web search in some degree the massivedata quantity and frequent update speed havebrought a completely new challenge as well.

3 PageRank Algorithm

PageRank is a global link analysis algorithm proposed by S. Brin and L. Page (Brin, S & Page, L 1998). It performs statistics to the URL link condition of whole Web, and calculates a weight, which is called as the PageRank value of this page, to every URL according to the factors such as link times, etc. This PageRank value is fixed, not changeable with the change of query keyword, which is different from the local link analysis algorithm HITS.

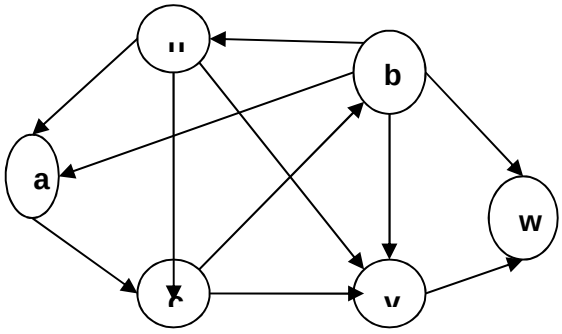


Figure 1 Directed link graph G

For example, in Figure 1, page u includes a hyperlink referring to page v, there exists $link(u,v)$. Here, the hyperlinks between pages compose a directed graph G. To a node composing directed graph G in every page, if and only if when u includes the hyperlink referring to page v, there exists directed edge (u,v) from u to v.

To node v, nodes b, c, u have contributions to the weight value of v, because these three nodes all exist directed edges to v. The more the directed edges referring to some node, the higher the node (page) quality is. The main shortcoming of this kind of algorithm is that only the link quantity is considered, which means all the links are equivalent, but whether the quality of source node itself is high or low is not considered. In fact, the high-quality page in Web always includes high-quality links, to the effect of linked document quality evaluation, the impact of the quality of source node is always high than the quantity. For example, the links appearing in Yahoo always have certain reference value, because Yahoo itself is a relatively authoritative Website, just as the papers issued in top publications always have higher academic value.

PageRank algorithm is in recursive form, its value relies on the linked times and the PageRank value of source link (Brin, S & Page, L 1998).

3.1 Simplified PageRank Algorithm

Simplified PageRank algorithm implements the basic recursive procedure of link times and source PageRank. Let the pages on Web as $1, 2, \ldots, m$, $N(i)$ is the amount of the extra-Website links of page i, $B(i)$ is the page set referring to page i. Assume Web is a strong connected graph (actually it is impossible, this problem will be discussed in the next section), then the PageRank value of page i can be expressed by:

$$r(i) = \sum_{j \in B(i)} \frac{r(j)}{N(j)}$$

The expression above can be written as $r = A^T r$, r is the vector of $m \times 1$, the arbitrary element in matrix A, which $a_{ij} = \frac{1}{N(i)}$. If page i refers to j, then $a_{ij} = 0$. Thus, vector r is the eigenvector of matrix A^T. Because Web is assumed to be strong connected, the eigenvalue of A^T is 1.

From the definition above, we can find that PageRank is accord with Random Surfer Model (Page, L, Brin, S &Motwani, R 1998). We can consider Random Surfer Model in this way: Assume a user visits Web page by means of

ISSN 2321-6905
January-2015

randomly clicking hyperlinks, moreover, he doesn't use "back" function and keeps continuous clicking. The PageRank of page i is essentially the probability of clicking page i in the process that a user browses the whole Web by means of random surfer. Motwani, R &Raghavan, P (1995) had made further research on RSM, these works can be also used to analyze the Web link attributes.

The computation of simplified PageRank algorithm can use iterative method, after several times of iteration, stop the iterative procedure when the PageRank value converge to the condition that deviation is small enough. For example, in Figure 2, the computation procedure is shown below:

1 Select arbitrary random vector s

1 $r = A^T \times s$

2 If $|r - s| < £$ ($£$ is the selected iterative threshold value), stop iteration. r is the PangeRank vector

3 $s = r$, back to step 2

Figure 2 shows the computed rank value of every node in a small graph structure by simplified PageRank algorithm. According to the RSM of PageRank, the sum of the PageRank value of every node is 1.

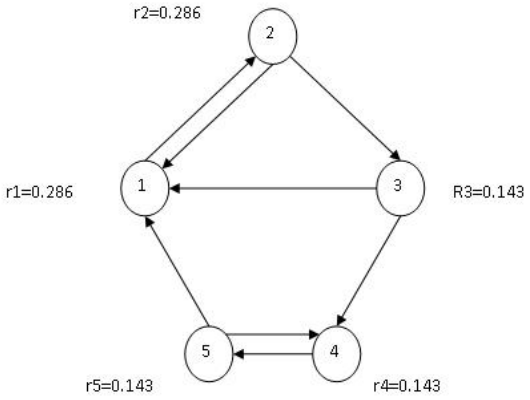


Figure 2 Simplified PageRank Algorithm

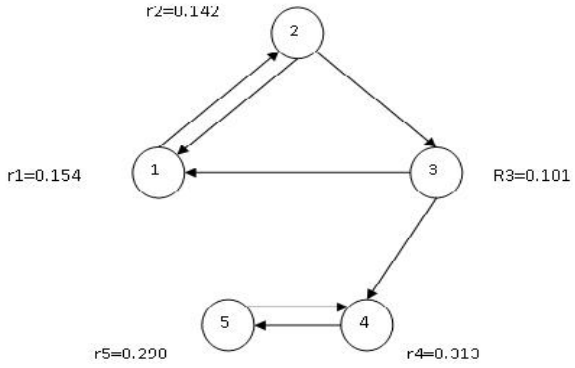
3.2 Improved PageRank Algorithm

Simplified PageRank algorithm is only suitable for the ideal strong connected environment, but in fact, Web is not a strong connected structure. Border, A, Kumar, R, &Maghoul, F's (2000) paper shows there are only 28% pages on Web are strong connected; 44% are one-way connected; and the remaining part forms Information Isolated Island, which is neither linked by, nor links to other page. To simplify PageRank algorithm, non-strong connected Web exists two inextricable problems, which are rank sink and rank leak. Rank sink refers to some local strong connected Web graph doesn't include the link referring to outside. Rank leak refers to the page that doesn't include any external hyperlink. Actually, it is a special case of rank sink when there is only one node in the strong connected graph. They will cause deviation generating when analyzing graph structure. For example, if we discard the link from 5 to 1 in Figure 2, nodes 4 and 5 will form rank sink situation. If we use RSM to simulate, we will fall into the dead circulation from 4 to 5 at last. Moreover, the rank values of 1, 2 and 3 tend to 0, and the nodes 4 and 5 will share the rank, which the total value is 1, of whole graph. If we remove 5 and its related links form figure 2, node 4 will become a leak node. Because once this node is visited, the rank procedure will stop here, thus, the rank values of all nodes will converge to 0. Therefore, Page and Brin (Brin, S & Page, L 1998) proposed two methods, one is discarding all the leak nodes which their outdegrees are 0, another one is introducing damping fact d ($0 < d < 1$) in simplified PageRank algorithm. The appearance of d makes PageRank contribute to not only the node which it links to, but also the other pages on Web. The expression of improved PageRank algorithm is shown below:

$$r(i) = d * \sum_{j \in B(i)} \frac{r(j)}{N(j)} + \frac{1-d}{m}$$

m is the total node amount of Web subgraph that Web Crawler visits. As we can see from the expression, the simplified PageRank algorithm is the special case when $d = 1$.

Figure 3 shows the computed PageRank value of every node after removing the hyperlink from 5 to 1 by improved algorithm. Every node has been adjusted by parameter d, which make their values all converge to a non 0 value.



ISSN 2321-6905
January-2015

Figure 3 Improved PageRank Algorithm

For example, in Figure 1, the PageRank value of each node is shown in the table below ($£ = 0.2$):

Table 1 PageRank of each node in Figure1

	Node a	*Node b*	*Node c*	*Node u*	*Node v*	**Node w**
PageRank	0.0 0 6 2 1 0	0.0 7 0 1 0 4	0.0 9 1 5 7 7	0.0 4 7 3 4 4	0.0 9 8 8 8 1	0.1 2 5 8 3 9

PageRank can use iterative algorithm to complete recursion. To the PageRank of each node in Figure 1, about 15 times iterations are needed. Generally, in actual computation, 100 times iterations are enough to converge (Haveliwala, H.T 1999).PageRank algorithm is currently applied by Google search engine, which provides high-quality Web retrieval service.

CONCLUSION

In this paper, through the research and analysis on classical link page ranking structure-based algorithms and their related improvements, we will propose an improved P-R algorithm based on categorization technology to navigate through the web pages for the effective retrieval of content from the web pages.

REFERENCES

[1] B. SakthiSaravanan., M.Tech#1, R.DheenadayaluM.Sc (Engg) #2 “Improving Efficiency and Security Based Data Sharing in Large Scale Network” International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 1, January 2013, ISSN: 2319-5967

[2] Mr. RupeshVaishnav “Attribute Based Signature Scheme For Attribute Based Encrypted Data In Cloud” International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 10, December 2012 ISSN: 2278-0181.

[3] M. Armbrust, A. Fox, R. Gri±th, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, A view of cloud computing," Communications of the ACM, vol. 53, pp. 50{58, 2010.

[4] D.Khader ,” Attribute Based Authentication Schemes,” PhD Dissertation University of Bath, 2009.

[5] M. D. Dikaiakos, D. Katsaros, P. Mehra, G. Pallis, and Athena Vakali, Cloud computing: Distributed internet computing for it and scientific research," IEEE Internet Computing, vol. 13, pp. 10 {13, 2009.

[6] A. Sahai and B. Waters. Fuzzy identity based encryption. In Eurocrypt 2005.

[7] D. Boneh and M. K. Franklin. Identity-based encryption from the weil pairing. In Proceedings of the 21st Annual International Cryptology Conference on Advances in Cryptology, pages 213–229. Springer Verlag, 2001.

[8] C. Cocks. An identity based encryption scheme based on quadratic residues. In IMA Int. Conf., pages 360–363, 2001.

[9] A. Shamir. Identity based cryptosystems and signature schemes. In Proceedings of CRYPTO 84 on Advances in Cryptology, pages 47–53. Springer Verlag New York, Inc., 1985.

[10] Matthew Pirretti , Patrick Traynor “Secure Attribute Based Systems”

Mr. J Prasanth Kumar is pursuing M.Tech Computer Science from ANDHRA LOYOLA INSTITUTE OF ENGINEERING AND TECHNOLOGY and completed B.Tech I.T in the year 2012 in ALIET, Vijayawada A.P. His areas of interests are Data Mining, Cloud Computing, Network Security and Web Tools