Formal Drift Framework for Higher Pagerank Scores

Saurabh Ranjan Srivastava[1], Girdhari Singh[2]
Department of Computer Science and Engineering
[1]Swami Keshvanand Institute of Technology Management & Gramothan, Jaipur
[2]Malviya National Institute of Technology, Jaipur
Email- [1]saurabh.ranjan.srivastava@gmail.com, [2]girdharisingh@rediffmail.com
Received 15 August 2015, received in revised form 15 September 2015, accepted 6 October 2015

Abstract: **Pagerank is a significant component of Google's ranking system of web documents [1]. Among the various off-page ranking parameters [2], Pagerank is the only confirmed and well-studied component. Therefore analysis of Pagerank is crucial for any web document to achieve higher search positions in the search results. This paper proposes a formal drift framework machine for structuring web applications capable of achieving higher Pagerank, crucial for better positions in the results of search engines. By exploiting the score computation mechanism of Pagerank algorithm, this framework expedites the Pagerank for every individual webpage linked to a web application towards a better score.**

Keywords: **Pagerank, on-page parameters, off-page parameters, HITS, inlink, outlink, InDegree, feedback.**

1. INTRODUCTION

In today's massive information intensive internet, providing most accurate and latest search results in minimum time is a major challenge for search engines. Due to this reason, the original search algorithm of every search engine is a deep hidden business secret [2]. Such algorithms are believed to work upon multiple parameters to rank web documents. The search ranking parameters for web documents can be classified into 2 major categories.

First category is of the on-page parameters. These are the factors that are present over the webpage itself like html tags, date of modification, version etc. The other category is the off-page parameters that are not present on the webpage to be ranked. Examples of such off-page factors are IP address of the webpage, hyperlinks pointing to that page and so on.

It is well known that the web is a massive collection of interrelated documents connected via hyperlinks [3-4]. Therefore, hyperlinks on web documents are a major factor of ranking [3]. Viewing the rapid change and diversity in the design of the content on the web, traditional dictionary based approaches of information retrieval are now obsolete [5-7]. Hence focus of the information retrieval research has shifted towards automated machine learning techniques. Automated search engines employ search ranking algorithms and keywords to return matches of varying quality and precision.

The search ranking results can be constantly improved by exploiting the hyperlink structure between the web documents [6]. Here we discuss two graph traversal algorithms, Pagerank [7] and HITS [8] for the purpose above stated. Both of these algorithms utilize the heuristic that webpages with higher number of incoming links are of higher importance compared to those with fewer incoming links.

2. RANKING ALGORITHMS

InDegree is the fundamental heuristic behind all hyperlink analysis based search ranking techniques that rank a web document by analyzing the incoming hyperlinks or the number of pages pointing to it [6]. Every recommendation by one webpage to another is a kind of vote that increases its popularity [3]. The 2 hyperlink based search ranking algorithms worth mentioning are HITS and Pagerank which we are discussing them ahead [9].

HITS

Hypertext Induced Topic Selection (HITS) [2] algorithm, proposed by Kleinberg, utilizes the InDegree heuristic for popularity of web documents. HITS ranks webpages by analyzing their Inlinks and Outlinks. Webpages referred by many hyperlinks are called *authorities* whereas webpages that point to many hyperlinks are called *hubs* [8-10]. The algorithm starts traversal with a root set of pages acting as a seed library of documents. This library increases by addition of the referee and referred pages in the root set. This method provides a small subgraph that is relatively focused on the query topic, with many relevant pages and strong authorities.

HITS separates hubs and authorities within a subgraph of relevant pages. For a given any set of web pages, and a specific query string, HITS restricts the analysis to the set of all pages containing the query string. HITS is applied on a subgraph after a search is done on the complete graph to define recursive relationship between webpages by use of hubs and authorities given as follows:

x = An authority is a page that many hubs link to

y = A hub is a page that links to many authorities

The scores for authority nodes x can be determined from the hub scores given as

$$\mathbf{x} = \mathbf{A}^T\mathbf{y}$$

Similarly the hub scores from the authority scores

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

Substituting into the equations we get the following :

$\mathbf{x} = \mathbf{A}^T\mathbf{A}\mathbf{x}$

$\mathbf{y} = \mathbf{A}\mathbf{A}^T\mathbf{y}$

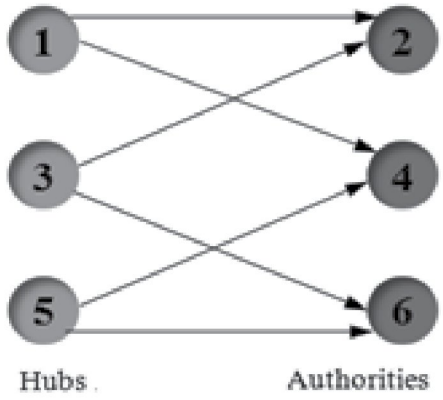

Figure-1: Simple web framework for HITS algorithm

Pagerank

A more democratic use of hyperlinks to evaluate the significance of webpages was presented in the Pagerank algorithm [1]. The PageRank algorithm involves computing the principal eigenvector of the Markov matrix representing the hyperlink structure of the web [1, 4, 11]. Based on the random surfer model, Pagerank covers the standing probability distribution of a random walk on the graph of the web [7, 12]. Hence, each step in the PageRank algorithm is of one of two types:

1. From the given states **s**, select a random outgoing link from the **s**, and follow that link to the destination page.

2. Choose a web page uniformly at random, and jump to it.

Due to its massive size, the PageRank vector is computed offline from the servers, during the preprocessing of the Web crawl, before any queries have been issued [9]. While PageRank computes the page ranks on the entire web graph, the HITS algorithm tries to distinguish between hubs and authorities within a subgraph of relevant pages.

We proceed our discussion on Pagerank scores of webpages from here onwards.

3. EFFECT OF LINK ARCHITECTURE ON PR SCORE [1, 2, 12]

The PageRank score of a webpage can be considered as a vote casted by other webpages for that webpage. A link to a page is considered as a vote of support. Mathematically, PageRank can be formulated as follows.

$$PR(A) = (1-d) + \{\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \cdots + \frac{PR(T_n)}{C(T_n)}\}$$

Here the values of various variables are as follows.

T_i = Webpage that links to webpage A

$C(T_i)$ = Number of outbound links on page T_i

n = Total number of webpages linked to webpage A.

As the PageRank forms a probability distribution over the given set of web pages, so the sum of all web pages PageRank will be 1.

Feedback Effect

Feedback is an obvious effect of the PageRank computation within internal site links, and is crucial to Google's evaluation of page significance within a site [6]. If the site had no incoming or outgoing links, the structure of the site would provide the same amount of feedback [9, 7].

The PR of each page depends on the PR of the pages pointing to it. But as the PR of the pointing pages has to be calculated iteratively, an iterative algorithm corresponding to the principle eigenvector of the normalized hyperlink matrix computes the Pagerank for a given set of pages [11]. It implies that the PR of a webpage can be calculated without knowing the final value of the PR of the other pages**.** For a simple web model of two pages each pointing to the other we have.

Figure-2: Two node model for Pagerank Computation

For each page with 1 outgoing link, the outgoing count is 1,

$C(1) = 1$ and $C(2) = 1$

For a 2 page model having 1 incoming and outgoing hyperlinks each, with an arbitrary PR= 1.0 and damping factor d = 0.85 we get:

$$PR(A) = (1-d) + d(PR_B)/1)$$
$$= 0.15 + 0.85 * 1 = 1$$
$$PR(B) = (1-d) + d(PR_A)/1)$$
$$= 0.15 + 0.85 * 1 = 1$$

The loss of Pagerank of a webpage caused due to outgoing links to external sites can be minimized by increasing the number of internal hyperlinks. This observation leads to analysis of following web link architectures.

In Figure-3 every page has 3 incoming and 3 outgoing hyperlinks. The double headed arrows in figure jointly represent an incoming and outgoing hyperlink.

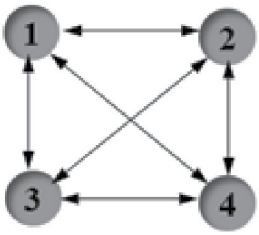

Figure-3: Extensive Interlinking

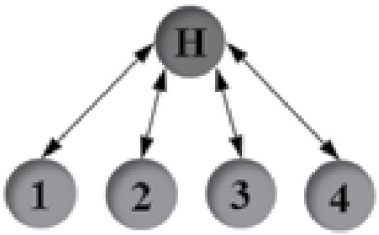

Figure-4: Centralized interlinking with feedback

But here also, the Pagerank scores for each webpage remain equal even after the 12th iteration. This implies that, on structuring the webpages in a fully connected extensive hyperlink topology, the earned Pagerank is evenly distributed among the connected webpages irrespective of the number of increased hyperlinks as visible in Table-1. Further, we have shown the centralized interlinking with feedback architecture. In this arrangement, every webpage is recursively connected with a central hub webpage, represented by H. The 'homepage' in many websites can be viewed as an example of the hub of this structure. It can be concluded from the Table-1, that the Pagerank of remaining pages is drifted towards the hub webpage, maximizing its Pagerank score. For this reason, generally the homepage of multiple websites achieves higher positions in search engine results due to higher individual Pagerank scores. The Pagerank scores for these 2 architectures after the 12th iteration are as follows:

NO. OF PAGES	EXTENSIVE INTERLINKING	CENTRALIZED INTERLINKING WITH FEEDBACK
Home page	1.0	2.182315
Page-1	1.0	0.704421
Page-2	1.0	0.704421
Page-3	1.0	0.704421
Page-4		0.704421

Table-1: Pagerank Values for 2 web frameworks

It can be observed that the Pagerank values of pages 1 to 4 have shifted, more appropriately, 'drifted' towards the homepage. Such drifting of Pagerank values can be utilized in an effective manner to improve the scores of specific target webpage. A similar approach is displayed in Figure-5 where the Pagerank scores for a target webpage given by T, in a group of other webpages are improved. For this, the incoming links for the target webpage T, internal to the website, are increased by connecting more internal webpages to it. The webpages are organized in a layered hierarchy from 1.1, 1.2, …, 1.x to m.1, m.2, … m.n, similar to that in Figure-4. Here m represents the number of layers in the website, while n is the number of webpages on the m^{th} level.

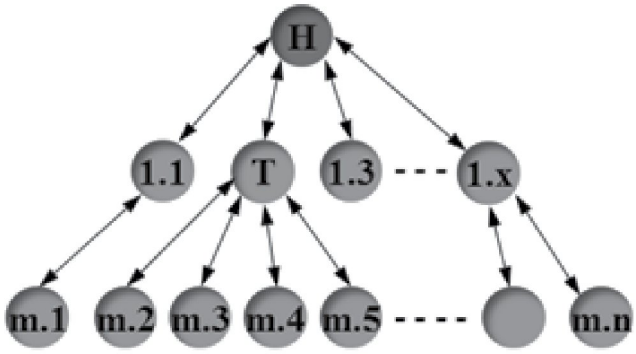

Figure-5: Framework for drifting Pagerank to target Pages

It should be noted that the page T is connected to lower layer webpages by both incoming and outgoing hyperlinks while the upper layered webpage connects to T by incoming and 1 outgoing link. This limiting of hyperlinks is crucial for minimizing the drifting of Pagerank value to higher level pages from the target webpage [10]. But at least a single outgoing link is also essential to maintain the Pagerank of webpages in upper level as well as providing a clear visibility of the target webpage to the search engine spider programs during web crawl process [7]. This linkage is essential to drift the Pagerank scores of lower layer webpages to page T. But connecting it only with leaf node webpages should not be a compulsion.

4. FORMAL DRIFT FRAMEWORK MACHINE

From the above cases, following patterns can be deduced for the computation and distribution of Pagerank scores.

1. Higher percentage of interlinking improves PageRank scores.

2. With Centralized interlinking, the PageRank of webpages in lower layers drifts to the webpages connected in higher levels of the web application. This means we are giving away less PageRank on outbound links.

3. On improving its Pagerank, a webpage may become a more preferred channel for the search engine spiders, to index web application by improving the overall ranking of the site in the search results.

By incorporating above stated patterns as guidelines, we propose a formal machine framework using the drifting nature of Pagerank score distribution [3, 6, 11], for designing of web applications.

Mathematically it can be represented as follows. Every web application can be considered as a set of webpages with certain Pagerank score. According to the random surfer model of web mining [1, 5, 12], a random user can click any webpage and start

accessing the other webpages connected via hyperlinks present on it. Similarly, he may stop and discontinue his surfing process at any random webpage.

Therefore, every webpage is considered as a start as well as a final state [5]. Still the category of main webpages, from where we expect maximum initialization, is marked as hub nodes (H). Hubs are connected to 2 major types of pages. First category is the general type of pages, present from level 1 to level m, organized in multiple layers. These pages have a general ratio of incoming and outgoing hyperlinks and may or may not have a high percentage of connectivity. Webpages displaying a higher ratio of connectivity can be considered as target as target nodes (T) whose Pagerank has to be improved.

The categories of webpages can be stated as follows.

H = Hub webpages initializing the web surfing

1.x ... m.n = Webpages distributed from level 1 to m.

T = Target webpages for which Pagerank has to improved

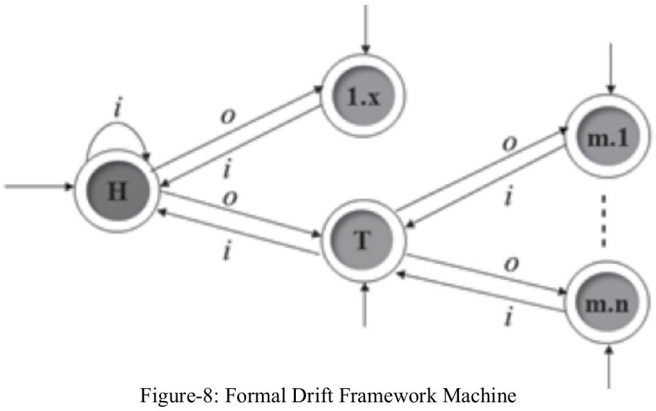

Figure-8: Formal Drift Framework Machine

This architecture presents a guideline for distributing the Pagerank scores via the centralized feedback model on the basis of hyperlink structure. It can be categorized into following generic layers with following components:

Level-H = Homepage / Index page.

The home page must present the prime content of the application and should attain feedback hyperlinks from all pages.

Level-1.x = Top classifier pages.

These pages should capture the secondary information about the domain addressed by the website. The site map, content description, categories of coverage can be also covered under this stage [10]. Finally textual recommendation for external resources can be included, but inclusion of any external hyperlink should be avoided to save the leakage of Pagerank to webpages of external resources.

Level-2 to m = The pattern mentioned in layer-1.x can be covered at these levels. All the dynamic pages should be included at this level as they keep on altering the overall Pagerank of the website.

Level-m+1 = The backend database and server logs, from where the application fetches data should be addressed at this level.

5. EXPERIMENTAL SETUP AND RESULT DISCUSSIONS

To prove our point, we consider an example of a group of webpages. Here webpage 2 is marked as the target page whose Pagerank score has to be improved.

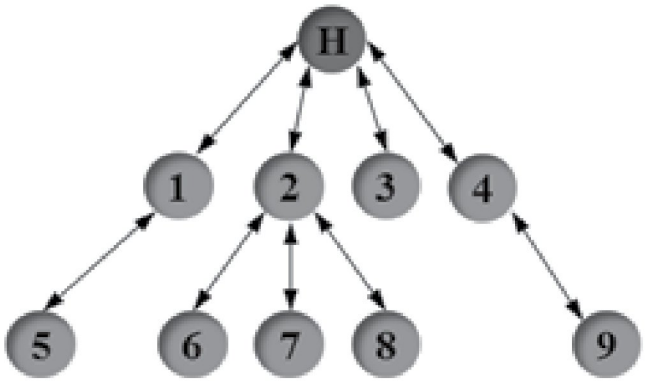

Figure-6: Drift Structure with 1 target and 1 hub

The comparison of Pagerank scores for 10 webpages in a centralized interlinking with feedback design and in a drift framework can be viewed in the tables 2 and 3.

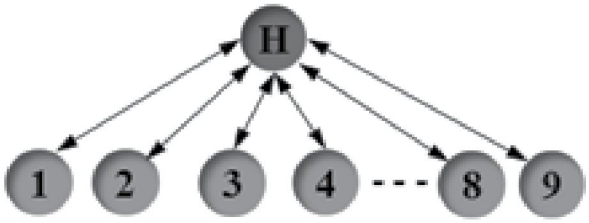

Figure-7: Centralized interlinking with feedback structure with 1 hub

The cumulative scores of the webpages can be compared to analyse the effect of additional internal hyperlinks to the target page from the evenly connected webpages of the centralized interlink structure.

To demonstrate our point, we established a PHP based Pagerank computing program for webpages. Interconnectivity of 9 webpages with the Hub webpage for the 2 architectures was provided as input to the PHP program.

Considering the centralized interlinking with feedback structure with 1 hub first, we find that all the pages attain a uniform Pagerank value of 0.244444. With successive iterations of the Pagerank mechanism, this value stabilizes to 0.649684 for each of the connected webpages.

But in the first iteration itself of the drift framework, the target webpage 2 attains a Pagerank score of 2.9125 which is even higher than the Pagerank of homepage which is 2.445. This

Pagerank score stabilizes towards a value of 1.594383 till the 12th iteration due to distribution of values among other webpages, but still attains a higher Pagerank score compared to other webpages.

The point to be noticed here is that in centralized feedback model, the Pagerank of hub or homepage has reduced from 7.8 to 4.152841 thus lowering its rank. While at the same time, the Pagerank of our drift framework has consistently evolved from 2.445 to 3.083592 by the 12th iteration.

Iterations	1	2	3	4	5	6	7	8	9	10	11	12
HUB	7.8	2.02	6.933	2.75695	6.306593	3.289396	5.854013	3.674089	5.527024	3.952029	5.290775	4.152841
Page1	0.244444	0.886667	0.340778	0.804783	0.410379	0.745623	0.460665	0.702879	0.496997	0.671997	0.523247	0.649684
Page2	0.244444	0.886667	0.340778	0.804783	0.410379	0.745623	0.460665	0.702879	0.496997	0.671997	0.523247	0.649684
Page3	0.244444	0.886667	0.340778	0.804783	0.410379	0.745623	0.460665	0.702879	0.496997	0.671997	0.523247	0.649684
Page4	0.244444	0.886667	0.340778	0.804783	0.410379	0.745623	0.460665	0.702879	0.496997	0.671997	0.523247	0.649684
Page 5	0.244444	0.886667	0.340778	0.804783	0.410379	0.745623	0.460665	0.702879	0.496997	0.671997	0.523247	0.649684
Page 6	0.244444	0.886667	0.340778	0.804783	0.410379	0.745623	0.460665	0.702879	0.496997	0.671997	0.523247	0.649684
Page 7	0.244444	0.886667	0.340778	0.804783	0.410379	0.745623	0.460665	0.702879	0.496997	0.671997	0.523247	0.649684
Page 8	0.244444	0.886667	0.340778	0.804783	0.410379	0.745623	0.460665	0.702879	0.496997	0.671997	0.523247	0.649684
Page 9	0.244444	0.886667	0.340778	0.804783	0.410379	0.745623	0.460665	0.702879	0.496997	0.671997	0.523247	0.649684

Table-2: Pagerank values for Centralized feedback model up to 12 iterations

Iterations	1	2	3	4	5	6	7	8	9	10	11	12
HUB	2.445	3.655188	2.634205	3.340544	2.727815	3.236702	2.800905	3.165991	2.854628	3.117684	2.893762	3.083592
Page1	1.2125	0.941563	1.229434	0.973324	1.165019	0.997806	1.133644	1.016875	1.114085	1.031047	1.100993	1.041411
Page2	2.9125	1.485563	1.834846	1.500436	1.775325	1.534097	1.725335	1.560241	1.696708	1.579924	1.677964	**1.594383**
Page3	0.3625	0.669563	0.926727	0.709768	0.859866	0.729661	0.837799	0.745192	0.822773	0.756608	0.812508	0.764925
Page4	1.2125	1.158313	1.492243	1.255709	1.526438	1.310786	1.516725	1.346214	1.49819	1.370428	1.481229	1.387492
Page 5	0.32	0.356125	0.310066	0.359004	0.315465	0.348053	0.319627	0.34272	0.322869	0.339394	0.325278	0.337169
Page 6	0.32	0.356125	0.310066	0.359004	0.315465	0.348053	0.319627	0.34272	0.322869	0.339394	0.325278	0.337169
Page 7	0.32	0.356125	0.310066	0.359004	0.315465	0.348053	0.319627	0.34272	0.322869	0.339394	0.325278	0.337169
Page 8	0.32	0.356125	0.310066	0.359004	0.315465	0.348053	0.319627	0.34272	0.322869	0.339394	0.325278	0.337169
Page 9	0.575	0.665313	0.642283	0.784203	0.683676	0.798736	0.707084	0.794608	0.722141	0.786731	0.732432	0.779522

Table-3: Pagerank values for Drift Framework model up to 12 iterations

6. COMPARISONS OF THE 2 ARCHITECTURES

As we have already discussed, a better Pagerank is a vital parameter of a higher position in the results of search engines. Any loss or fluctuation in its iterative computation may lead to depreciation of position in search results.

As evident from the comparisons of the Pagerank scores above, we can easily conclude that the drift framework helps to attain a good Pagerank for a specially targeted webpage. Such targeted webpages may contain the newly added or latest content of the website, that the webmasters wish to display on priority. In such cases, this framework can be used to highlight a specific webpage, without heavily disturbing the Pagerank of the homepage. A graphical comparison of the 2 approaches is given ahead. The Pagerank scores of the hub and 9 webpages are represented against iteration counts in space graphs.

Each line in the 3-dimensional space represents its individual progress of Pagerank values through the 12 iterations.

As visible in figure-8, for the centralized interlinking architecture, the Pagerank scores of all webpages transform in parallel. Here the hub webpage, which generally is the index or homepage, attains the highest Pagerank from start to end.

Contrary to this, each webpage in figure-9 attains a different range of scores in the graph space by using the drift framework architecture.

The target webpage, page 2 initiates from a score of 2.9125 and stabilizes to 1.594383 still with a second highest Pagerank score in the group. Targeting one or more webpages for Pagerank score improvement also reduces the fluctuation in the hub webpages of the group.

This is evident from the performance of hub webpage in figure-9 compared to that in figure-8 as the drifted Pagerank minimizes the changes in the scores of the webpages in higher levels of the group

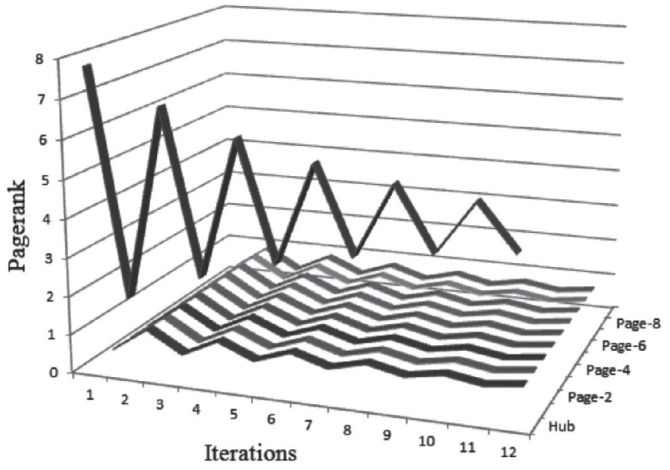

Figure-8: Transformation of Pagerank scores for webpages in Centralized interlinking with feedback structure

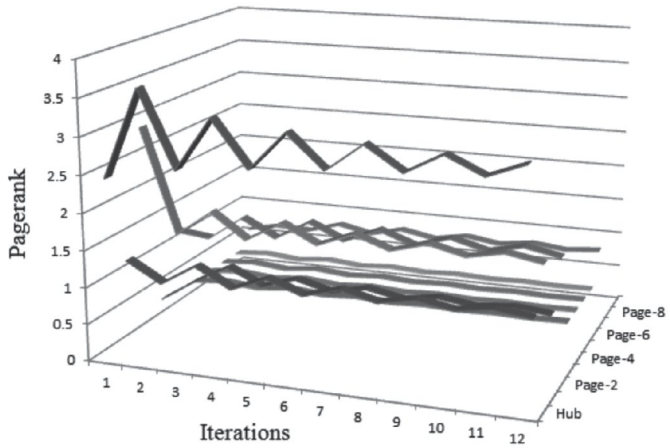

Figure-9: Transformation of Pagerank scores for webpages in Drift Framework

7. CONCLUSIONS AND FUTURE WORK

To capture best Pagerank by implementing the 'Formal Drift Framework Machine' model we propose following practices for web application design.

1. Organize the web application into layers of webpages according to their content relevance into categories with each category a different page of its own. In case more than 1 layer is present below the webpage T, it implies that the Pagerank of lower layered webpages will also drift towards page T.

   Mathematically we can conclude these conditions for webpage T as follows:

   $$\text{Hyperlink (T): } \{ \quad \text{Outgoing links} \geqslant 1$$
   $$\text{Incoming links} \leqslant n$$

2. Provide limited connectivity of the Hub pages such as homepage with feedback links to the remaining web model.

   This enables a single channel flow of Pagerank towards the top pages and minimizes Pagerank drainage to lower level pages.

3. Avoid structures such as sitemap that significantly lower the PageRank of other pages. This eventually reduces the Pagerank of pages in upper layers.
4. Interlink pages of each category together along with the category index page at level-1.x tier by use of a navigational menu.
5. Maximize the feedback hyperlink connectivity of internal webpages to the target webpage, to improve its score, and in turn enhancing the score of higher level pages also.

   The proposed drift architecture can be further tested for more web applications of higher complexity. Development of ontological libraries by keeping view on this design should lead to improved search rankings of the target web applications having better organization of data.

REFERENCES

[1]. L. Page, S. Brin, R. Motwani, and T. Winograd; The PageRank citation ranking: Bringing order to the web; 1998; Stanford Digital Libraries.

[2]. Golliher Sean A.; Search Engine Ranking Variables and Algorithms; 2008; SEMJ.Org Volume 1, Supplemental Issue, August 2008; Pages 15-19

[3]. Henzinger M.; Link Analysis in Web Information Retrieval; 2000; Bulletin of Technical Committee on Data Engineering; IEEE Computer Society; September 2000; Volume-23 Number-3; Pages 3–8.

[4]. Selvan Mercy P., ChandraSekar A., Dharshin Priya A.; Survey on Web Page Ranking Algorithms; 2012; International Journal of Computer Applications (IJCA); ISSN: 0975-8887; Volume 41– No.19, March 2012, , Page 1-7

[5]. M. Richardson, P. Domingos.; The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank; 2002; Advances in Neural Information Processing Systems 14, volume 14.MIT Press, Cambridge, MA

[6]. Smitha L., Fatima S S.; Query Independent Time Dependent Page Ranking Algorithm for Web Information Retrieval; 2012; Proceedings on International Conference and workshop of Emerging Trends in Technology (ICWET 2012); ICWET 2012 - Number 9; Pages 6-10

[7]. Brin S., Page L.; The Anatomy of a Large-Scale Hypertextual Web Search Engine; 1998; Computer Networks and ISDN Systems; Volume 30 Issue 1-7, April 1, 1998; Pages 107-117

[8]. Prajapati R.; A Survey Paper on Hyperlink-Induced Topic Search (HITS) Algorithms for Web Mining; 2012; International Journal of Engineering Research and Technology (IJERT); ISSN: 2278-0181; Volume-1 Issue-2, April – 2012, Page 13-20

[9]. Tyagi N., Sharma S.; Comparative study of various Page Ranking Algorithms in Web Structure Mining; 2012, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-1, June 2012, Page 14-19

[10]. Hegde Mamta M., Phatak M.V.; Developing an approach for hyperlink analysis with noise reduction using Web Structure Mining; 2012; International Journal of Advanced Research in Computer Engineering & Technology; ISSN: 2278-1323; Volume 1, Issue 3, May2012, Page 68-72

[11]. Pretto L.; A Theoretical Analysis of Google's PageRank; 2002; Chapter String Processing and Information Retrieval-SPRINGER; Volume 2476; Series Lecture Notes in Computer Science; September 2002; Pages 131-144;

[12]. Agarwal A., Chakravarti S.; Learning Random Walks to Rank Nodes in Graphs; 2012; Proceedings of the 24th International Conference on Machine Learning (ICML); ACM New York, NY, USA ©2007; ISBN: 978-1-59593-793-3; Pages 9-16

◆ ◆ ◆