Web Page Ranking Based on Text Content of Linked Pages

P. C. Saxena, J. P. Gupta, Namita Gupta

Abstract—**In the traditional Information retrieval system, ranking of the documents is done based on the relevance of the document w.r.t. to the searched query. Relevance of the document is computed entirely based on text content of the document. But due to large number of web pages, searching on the web results in large set of web pages retrieved as a result. Effective ranking of these resultant pages is required in order of their relevance to the searched query. The link information of these web pages plays an important role while ranking them. Different link Analysis ranking algorithms are suggested which compute the ranking of web pages like Kleinberg's HITS algorithm, Lempel and Moran's SALSA algorithm, BFS algorithm and many improved modified algorithms. All these link analysis ranking algorithms (LAR) have their limitations that show that any ranking algorithm cannot rely solely on link information, but must also examine the text content of linked sites to prevent the difficulties observed by existing link analysis ranking algorithms. In this paper, we study the ranking scores of pages computed through different link analysis ranking algorithms and proposed a new ranking approach based on the content analysis of the link pages while computing the rank score of the target web page.**

Index Terms—Backward links, Forward links, Information Retrieval, Link Structure Analysis, Web page ranking.

I. Introduction

To manage the rapidly growing size of World Wide Web and to retrieve only related Web pages when given a searched query, current Information retrieval approaches need to be modified to meet these challenges. Presently, while doing query based searching, the search engines return a list of web pages containing both related and unrelated pages and sometimes showing higher ranking to the unrelated pages as compared to relevant pages. These search **engines** use one of the following approaches to organize search and analyze information on the web. In the first approach [10], the search engine selects the terms for indexing a web page by analyzing the frequency of the words (after filtering out common or meaningless words) appearing in the entire or a part of the target web page. The second method [1], [6], [9], [14], [23] uses the structure of the links appearing between pages to identify pages that are often referenced by other pages. Analyzing the density, direction and clustering of links, such method is capable of identifying the pages that are likely to contain valuable information. Another method [4], [11], [20] analyzes the content of the pages linked to or from page of interest. They analyze the similarity of the word usage at the different link distance from the page of interest and demonstrate that structure of words used by the linked pages enables more efficient indexing and search. Anchor text [15] of a hyperlink is considered to describe its target page and so target pages can be replaced by their corresponding anchor text.

But the nature of the Web search environment is such that the retrieval approaches based on single sources of evidence suffer from weaknesses that can hurt the retrieval performance. For example, content-based Information Retrieval approach does not consider the pages link by the page while ranking the page and hence affect the quality of web documents, while link-based approaches [2], [9], [14] can suffer from incomplete or noisy link topology. This inadequacy of singular Web Information Retrieval approaches make a strong argument for combining multiple sources of evidence as a potentially advantageous retrieval strategy for Web Information Retrieval.

Our system is built on an idea that to rank relevant pages higher in the retrieved document set, an analysis of both page’s text content and links information is required. Our approach is based on the assumption that the effective weight of a term in a page is computed by adding the weight of a term in the current page and additional weight of the term in the linked pages. In rest of the paper, we study various link analysis ranking algorithms and their limitations and show the comparative analysis of the ranking scores obtained through these approaches with our new suggested ranking approach.

II. Background And Previous Work

In this section, we discuss the necessary background for the rest of the paper. Also we review the various existing ranking algorithms and their limitations which is then used as a base for our new ranking approach.

Preliminaries

All the link analysis ranking algorithms [2] use the in-links (backward links pointing to a page) and out-links (forward links pointed by the page) of a web page to score the retrieved web pages. Initially a search engine is used to retrieve a set of web pages relevant to the given searched query. This creates a Root set. Then this Root Set is expanded to obtain a larger Base Set of Web pages by adding those pages which are pointing to the pages (backward links) of the original Root Set

P.C. Saxena is with the Department of School of Computer and System Sciences, JNU, New Delhi, India (e-mail : pcs@jnuniv.ernet.in).

J.P. Gupta is with the Jaypee Institute Of Information Technology, JIIT University, Noida, Uttar Pradesh, India (e-mail : jp.gupta@jiit.ac.in).

Namita Gupta is with the Maharaja Agrasen Institute of Technology, GGSIPU, New Delhi, India, IACSIT membership No. 80333020
(e-mail: nimi_goyal@yahoo.com).

and the pages which are pointed to by the pages (forward links) of the original Root Set. Next, a hyperlink directed graph *G = (V,E)* is constructed from the Base set with the web pages defining the nodes *1, . . . , n,* and the links between the web pages defining the edges in the graph. This graph is described by an $n \times n$ adjacency matrix *A*, where $a_{ij} = 1$ if there is a link from page *i* to page *j* and $a_{ij} = 0$ otherwise. The vector $B(i) = \{j{:}a_{ji}=1\}$ represents the set of nodes that point to node *i* (backward links) and the vector $F(i) = \{j{:}a_{ij}=1\}$ represents the set of nodes that are pointed to by node *i* (forward links). All the link-based ranking algorithms are based on the idea that a web page serves two purposes: to provide information on a topic, and to provide links to other pages giving information on a topic. This gives rise to two ways of categorizing a web page. First, a web page to be an authority on a topic if it provides good information about the topic and is defined as authority node in graph *G* having nonzero in-degree . Second, a web page to be a hub if it provides links to good authorities on the topic and is defined as hub node in graph G having nonzero out-degree.

Let, *a* denote the set of authority nodes, *h* denotes the set of hub nodes, $G_a = (a,E_a)$ denotes the undirected authority graph on the set of authorities *a* having an edge between the authorities *i* and *j*, if $B(i) \cap B(j) \neq \phi$.

Previous Work

In recent years, a number of papers [7], [8], [11], [12], [13], [16], [18], [19], [20] have considered the use of hypertext links to determine the ranking score of different web pages. In particular, these papers consider the extent to which hypertext links between World Wide Web documents can be used to determine the relative authority values of these documents for various search queries. Also the link structures are used for categorizing pages and clustering them [5], [16]. Here in this paper we discuss some previous link analysis ranking algorithms [1], [2], [19] which we will consider while comparing our new ranking approach.

HITS (Hyperlink Induced Topic Distillation)

HITS algorithm is based on the idea that there is a mutual reinforcing relationship between the authorities and hubs. A good hub points to good authorities and a good authority is pointed to by good hubs. In order to quantify the quality of a page as a hub and an authority, Kleinberg associated with every page a hub and an authority weight. It uses an iterative algorithm for computing the hub and authority weights. Initially all authority and hub weights are set to 1.At each iteration, the authority and hub weight of a node is computed. Thus, for some node *i*,

$$a_i = \sum_{j \in B(i)} h_j \quad and \quad h_j = \sum_{i \in F(j)} a_i \qquad (1)$$

The algorithm iterates until the vectors converges. HITS consider the whole graph, taking into account the structure of the graph around the node to compute its hub and authority scores.The *ARC* system, described in [20], augments Kleinberg's link-structure analysis by considering also the anchor text, the text which surrounds the hyperlink in the pointing page. ARC computes a distance-2 neighborhood graph and weights edges. The weight of each edge is based on the match between the query terms and the text surrounding the hyperlink in the source document. The reasoning behind this is that many times, the pointing page describes the destination page's contents around the hyperlink, and thus the authority conferred by the links can be better assessed. Henzinger and Brahat [14], have also studied Kleinberg's approach and have proposed improvements to it. The connectivity analysis has been shown to be useful in identifying high quality pages within a topic specific graph of hyperlinked documents. The essence of their approach is to augment a previous connectivity analysis based algorithm with content analysis. The results of a user evaluation are reported that show an improvement of precision at 10 documents by at least 45% over pure connectivity analysis.

SALSA (Stochastic Approach for Link Structure Analysis)

Like Kleinberg's HITS algorithm, SALSA [3], [18] starts with a similarly constructed Base Set. It performs a two-step random walk on the bipartite hub and authority graph, alternating between the hub and authority sides. The random walk starts from some authority node selected uniformly at random. When at a node on the authority side, an incoming link is uniformly selected at random and moves to a hub node on the hub side. Each hub divides its weight equally among the authorities to which it points and the authority weight of a node is computed by summing up the weight of the hubs that point to it. Similarly, when at a node on the hub side, the algorithm selects one of the outgoing links uniformly at random and moves to an authority. Each authority divides its weight equally among the hubs that point to it and the hub weight of a node is computed by summing up the weight of the authorities that it point to.

Thus, for some node *i*,

$$a_i = \sum_{j \epsilon B(i)} \frac{1}{|F(j)|} h_j \quad and \quad h_i = \sum_{j \epsilon F(i)} \frac{1}{|B(j)|} a_j \qquad (2)$$

A significant advantage of SALSA is that the weightings can be computed explicitly without the iterative process. Hence it is a quicker method of calculating the weightings and can be generalized to accommodate non-uniform initial weights.

pSALSA (popularity SALSA)

It is a simplified version of SALSA and also performs a two-step random walk on the bipartite hub and authority graph, alternating between the hub and authority sides. But the starting point for the random walk is chosen with probability proportional to the "popularity" of the node, that is, the number of links that point to this node.

HUBAVG (Hub-Averaging)

The HUBAVG algorithm is a hybrid of HITS and SALSA algorithm. The idea behind the algorithm is that a good hub should point only to good authorities, rather than to both good and bad authorities. It updates the authority weight of a node like the HITS algorithm, but it sets the hub weight of node *i* to the average authority weight of the authorities pointed to by hub *i*. Likewise, it updates the hub weight of a node as in the SALSA algorithm, but it sets the authority weight of node *i* to the average hub weight of the hubs pointed to authority *i*.

Thus, for some node *i*,

$$a_i = \sum_{j \epsilon B(i)} h_j \quad and \quad h_i = \frac{1}{|F(i)|} \sum_{j \epsilon F(i)} a_j \qquad (3)$$

This approach has the limitation. Consider two hubs pointing to an equal number of equally good authorities. The two hubs are identical until one puts one more link to a low quality authority. The average sum of the authorities it points to sinks, and it gets penalized in weight. This limitation is removed by using Authority Threshold Algorithm.

HThresh (Hub Threshold)

The algorithm is similar to HITS, but to compute the authority weight of the *i*th page, it does not consider all hubs that point to page *i* but only those whose hub weight is at least the average hub weight over all the hubs that point to page *i*, computed using the current hub weights for the nodes. This overcomes the drawback of pSALSA by assigning lower authority weight to a site which points to large number of poor hubs.

AThresh (Authority-Threshold)

This algorithm is similar to HITS, but to compute the hub weight of the *i*th page, it does not consider all authorities that are pointed by page *i* but only those authorities which are among the top K authorities, judging by current authority values. Hence, for a site to be a good hub, it must point to some of the best authorities. The algorithm is based on the fact that, in most web searches, a user only visits the top few authorities.

Thus, for some node *i*,

$$a_i = \sum_{j \epsilon B(i)} h_j \quad and \quad h_i = \sum_{j \epsilon F_k(i)} a_j \qquad (4)$$

FThresh (Full-Threshold algorithm)

This algorithm makes both the Hub-Threshold and Authority-Threshold modifications to Kleinberg's HITS Algorithm.

BFS (Breadth-First-Search Algorithm)

Breadth-First-search algorithm ranks the nodes according to their reach ability i.e., the number of nodes reachable from each node. The algorithm starts from node *i*, and visits its neighbors in BFS order, alternating between backward and forward steps. Every time one link is moved further from the starting node *i*, the weight factor of the node is updated accordingly. The algorithm stops either when *n* links have been traversed, or when the nodes that can be reached from node *i* are exhausted.

The weight of node i is determined as:

$$a_i = |B(i)| + \frac{1}{2}|BF(i)| + \frac{1}{2^2}|BFB(i)| + \ldots + \frac{1}{2^{2n-1}}|(BF)^n(i)|$$
$$= \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k w^k [j,i] \qquad (5)$$

where, $(BF)^n(i)$ denote the set of nodes that can be reached from *i* by following a $(BF)^n$ path.

Following observations have been made [1], [2], [3], [18] about the different link analysis ranking algorithms:

1) Kleinberg algorithm is biased towards tightly-knit communities (TKC) and ranked set of small highly interconnected sites higher than those of large set of interconnected sites which is having hub pointing to a smaller part of the authorities.

Inappropriate zero weights can be seen in HITS regardless of the output's dependence on or independence of the initial vector. In multi-topic collections, the principal community of authorities found by the Kleinberg approach tends to pertain to only one of the topics in the collection.

2) For both HITS and SALSA, there are some graphs that give rise to repeated eigenvalues. The output of such graphs is sensitive to the initial vector chosen.
3) pSALSA algorithm place greater importance on the in-degree of a node when determining the authority weight of a node and favors various authorities from different communities. The algorithm is local in nature and the authority weight assigned to a node depends only on the links that point to the node. But counting the in-degree as the authority weight is sometimes imperfect as it sometimes results in pages belonging to unrelated community ranked higher than the pages belonging to related community.
4) Hub-average algorithm also favors nodes with high in-degree. It overcomes the shortcoming of the HITS algorithm of a hub getting a high weight when it points to numerous low-quality authorities. So to achieve a high weight a hub should link good authorities. But the limitation of the algorithm is that a hub is scored low as compared to a hub pointing to equal number of equally good authorities if an additional link of low quality authority is added to it.
5) Threshold algorithms (AThresh, HThresh, FThresh) eliminate unrelated hubs when computing authorities and hence tries to remove the TKC effect as seen in HITS algorithm. The results obtained from threshold algorithms are 80% similar to HITS algorithm.
6) BFS algorithm exhibits best performance among all LAR algorithms. BFS is not sensitive to tightly-knit communities as the weight of a node in the BFS algorithm depends on the number of neighbors that are reachable from that node. It also avoids strong topic drift.

Apart from the above mentioned link analysis ranking algorithms which we have used in our paper there are some other ranking algorithms also which helped us in designing our new ranking algorithm and so are discussed below.

Related Work

Carrier and Kazman [5] proposed a ranking measure on WWW pages, for the goal of re-ordering search results. The rank of a page in their model is equal to the sum of its in-degree and its out-degree, thus, it makes use of a "directionless" version of the WWW link structure.

Wang and Kitsuregawa [22] proposed an improved clustering algorithm to cluster the web search results by computing the similarity between the documents using the keywords appearing in the anchor text of a hyperlink in a web page and the link information i.e., the number of out-links and the in-links common to two documents under consideration.

Eiron and McCurley [15] showed the relation of an anchor text to a document and how the documents retrieved by anchor text techniques improve the quality of web text search than documents retrieved by content indexing. Their study revealed that anchor text is less ambiguous than other types of texts like title of document which are typically longer than individual anchor text and thus they resembles real-world queries in terms of its term distribution and length. Also anchor text provides better indication of the summarization of the page in different contexts, by different people, than that afforded by a single title which is authored by one author.

Yang [11] proposed a Fusion method to remove the inadequacies of singular web IR approaches. His experimental results showed that on combining various content-based (VSM) and link–based (HITS) systems, the optimum performance level of one method can be raised by combining it with a reasonably effective method of a different kind. His analysis of results suggested that index source, query length, and host definition are the most influential system parameters for retrieval performance.

Westerveld et. al. [21] suggested an empty page finding task to retrieve the web pages based on information about the document's content along with its in-links, URLs and anchors. They characterize page URL in four categories- Root, SubRoot, Path and file. Their results show that each combination of content or anchor and another source of information outperform the content or anchor run. The proper combination of URL and in-link information (i.e., without the independence assumption) performs better than the two separate priors. It is observed that URL information gives the best prior information. Adding in-links yields marginal improvement. From the above study, it is observed that ranking of a web page is highly influenced by the following factors [15], [19], [21] *a*. Text content of a web page ***b***. Anchor text of the hyperlinks in the page *c*. In-links to a page and out-links from a page ***d***. Web page URL.

It is also observed that considering each factor individually does not retrieve good quality web pages. Content-based IR approaches have difficulty dealing with the diversity in vocabulary and quality of web documents, while link-based approaches suffer from incomplete or noisy link topology. This inadequacy of singular web IR approaches focuses the researchers to modify the existing conventional IR methods and to propose new approaches which use both page content and link information of web pages [11] to solve IR problems. In the next section we proposed a method which is based on both content information of a target page and information about its link pages to rank it. While ranking a web page, our method also considers the text content of the hyperlinked pages to reduce the error due to noisy links and also to prevent the "topic drift" problem.

The remainder of the paper is organized as follows. Section 3 discusses the nature of different types of web pages available on WWW and their preference order. Section 4 discusses our proposed ranking system. Section 5 shows the experimental results obtained from the new algorithm. Conclusions are mentioned in section 6.

III. Nature Of Web Pages

Web pages of different types are retrieved as a result of searched query from the WWW. The nature of information available in these pages varies. There are pages having no forward links and discusses about the relevant topic. There are also pages which are index pages having hyperlinks only without any description on the searched query topic. Sometimes some pages are retrieved which are not relevant to the topic. All the possible kinds of web pages are listed in the table 1 given below.

Category	Web page discussing on related topic	Web page having Forward links on related/similar topic	Web page having Back links on related/similar topic
1	Y	Y	Y
2	Y	Y	N
3	Y	N	Y
4	Y	N	N
5	N	Y	Y
6	N	Y	N
7	N	N	Y
8	N	N	N

Table1 List the different categories of web pages.

Let us discuss by an example the different nature of possible information contained in a web page.

1) A web page discussing topic related to searched query. For example, retrieved a web page on "text mining" for the searched query "text mining".
2) A web page containing Forward links on same topic of searched query. For example, retrieved a web page containing forward links to pages discussing the topic "text mining" for the searched query "text mining".
3) A web page containing Forward links on related topics of searched query. For example, retrieved a web page containing forward links to pages discussing the topic "text mining" for the searched query "mining".
4) A web page containing Forward links on unrelated topics of searched query. For example, retrieved a web page containing forward links to pages discussing the topics like "Spanning tree protocol" for the searched query "Spanning tree" whereas the user is interested in Spanning tree graph.

A web page belonging to category 1 is most relevant to the searched query and should be given highest ranking score among all the categories where as page belonging to category 8 is least relevant to the searched query and should be assigned lowest ranking score. In the next section we proposed our new approach for ranking the retrieved web pages which is designed considering the different information available in the web pages as discussed in this section.

IV. Proposed Method

In our study, we propose a method to compute the relevance of a page to a searched query based not only on the information contained in its textual content but also by computing the relevance of the linked pages to the current page w.r.t. to the given searched criteria. The proposed algorithm represents each page as a vector of terms using Vector Space Model technique (VSM). VSM estimates the relevance of each term to the page [17] using the term frequency information to generate weights for all the terms in a document and represents the documents as term frequency weight vectors, so that document j is represented by the vector

$(w_{ij}) = 1 ..... m$

where, m is the total number of unique terms appearing in the document.

Different methods are used to calculate the weight of a term. In the proposed method, we use Term Frequency (TF) Weighting approach to compute the weight of the term in each page. The weight of an *i*th term using TF weighting is

$$w_i = \frac{tf_i}{T} \qquad (6)$$

Where tf_i is the number of times the ith term appears in the document

T is the maximum frequency of any term in current page p

Page is ranked higher if it contains functional links (i.e., links to pages related to the same topic). To differentiate the forward links as functional or navigational links, the content of the forward link pages is considered and if it is related to the same topic then it is considered while computing the ranking of the target page by the ratio proportional to its relevance to the given topic. The idea behind is that if there are two pages having same number of forward links. Let first page is linking to a page which is also discussing the searched topic and second page is linking to a page which is not related to the searched topic, then in this case first page should be ranked higher than the second irrespective both have same number of out-links.

Hence, the additional weight of *i*th term in current page *p* due to forward links is computed as :

$$aw_i = \frac{1}{H}\left(\sum_{j=1}^{H} L_{ji}\right), \quad L_{ji} = \frac{tf_{ji}}{T} \qquad (7)$$

where, H is the number of pages linked by page p

tf_{ji} is the number of times the ith term appears in the jth document

T is the maximum frequency of any term in jth page

Likewise, a page is ranked higher if it is pointed to by pages that are also related to the same topic. This will remove the problem of assigning high rank score to a page due to large no. of inlinks although some of these inlinks not related to searched topic, but where a link is added to a page just to improve its ranking as seen in pSalsa LAR algorithm. Based on this concept, higher score can be assigned to a page with few backward links but having functional links in comparison to a page having large number of non-functional navigational backward links.

Hence, the additional weight of *i*th term in current page *p* due to backward links is computed as:

$$aw'_i = \frac{h'}{H'} \qquad (8)$$

where, h' is the total number of pages pointing to page p having tf_{ji} >= average term frequency in page j

H' is the number of pages pointing to page p

The effective weight of the ith term in page p is thus given as:

$$ew_i = w_i + aw_i + aw'_i = \frac{tf_i}{T} + \frac{1}{H}\sum_{j=1}^{H} L_{ji} + \frac{h'}{H'} \qquad (9)$$

Similarly, we can calculate the effective weight of each word in a page and stored it in the inverted_word_document table [17] against the corresponding word with the page information in its posting_list. Whenever, a search query is given, for each search query term, inverted_word_document table is searched to retrieve documents list against query term from the table.

Here we analyze the similarity of the word usage at single level link distance from the page of interest and demonstrate that information about the linked pages enables more efficient indexing and search. Sample data is collected and is experimented using the above proposed algorithm. In the next section we discuss the results obtained during the testing followed by the conclusion derived from these results.

V. Experimental DataSet

In order to test the effectiveness of our proposed algorithm, we use the same Base dataset as used by Borodin et al. in [1]. We apply our method on the same queries i.e., *abortion, computaional geometry, computational complexity, gun control, net censorship and genetic* using the same Base dataset and then compare the results. The results are recorded in table as shown in **Appendix.**

We implement our method in Linux using bash scripting. To collect the Backward and forward link pages of the Root set, we use *wget* command available in Linux. The results of eight link analysis ranking algorithms on the Root Set (*HITS, pSALSA, SALSA, HubAvg, AThresh, HThresh , FThresh, BFS*) are collected from the site : http://www.cs.toronto.edu/~tsap and are used in analyzing the accuracy of the ranking scores computed by our ranking algorithm. Also the ranking score of the recent modified web pages obtained from PageRank, Alexa Rank, AltaVista results, and AllTheWeb results are recorded for verifying our results. These scores are obtained using the links http://www.mywebpagerank.com/ and http://findmypagerank.com/index.php.The comparative analysis of the ranking scores of ten web pages in each category i.e., *abortion, computaional geometry, computational complexity* is listed in Appendix and the results inferred from the analysis are discussed below.

Results

The proposed method considers the page content of backward link pages, forward link pages and the content of the target page to compute the rank score of the target page. The proposed algorithm reduces the limitations of the other link analysis ranking algorithms by differentiating between navigational and functional links. It is also based on the concept that only good hubs are considered in computing the ranking of the target page and only good authorities contribute in computing the final ranking of the target page. A hub is considered good if it points to pages which are related to given same topic; similarly a good authority is the one which is pointed to by pages which are discussing the same given topic. This is clearly depicted in the results obtained by implementing the proposed algorithm for different queries on the base dataset as shown in Appendix.

The ranking of web pages computed by our proposed algorithm is comparable to the ranking score obtained by other LAR algorithms. Slight variations in the ranking of the web pages are due to error in retrieving some of the backward links and forward links of some root web pages. The reasons for this error can be modifications made in the web page or server containing the target page down at the time of searching of the web page.

For query "*Computational Geometry*", Page P-10 is assigned zero ranking score by all the LAR algorithms as listed in the table 1.2 in Appendix, because page P-10 belongs to category 6 having no backward link (poor hub) and only two forward links. The page itself doesn't contain text related to the topic "Computational Geometry" but contain only forward links related to the target topic. Since all the LAR algorithms discussed above are influenced by the in-degree of a web page while computing the ranking score and hence assigned zero rank score to P-10. While our ranking algorithm consider all the three factors (as listed in table 1) to compute the ranking score of a web page and hence assigned non-zero small rank score to P-10 as it has forward links which are linking to pages related to target topic. Table 1.2 and Table 1.5 show conflicts in the ranking score of web pages P-21, P-27 and P-31. Only PageRank algorithm shows zero ranking score to pages P-21, P-27 and P-31 while other ranking algorithms (Alexa Rank, AltaVista results, and AllTheWeb) computes non-zero rank score for these pages which are similar to the results obtained by our algorithm. Also as shown in table 1.5, page P-10 is having non-zero rank score which is similar to our results and hence strengthens the accuracy of the results obtained by our algorithm. The ranking score of other web pages are comparable with the scores obtained by our ranking method as shown in table 1.4.

The results of the "*abortion*" query as shown in table 2.2 of Appendix shows zero rank score for many web pages as computed by different LAR algorithms. Web page P-81 has zero ranking score in many LAR algorithms (Kleinberg, HubAvg, AThesh, FThresh) or very small rank score in others (pSalsa, Salsa, HTresh, PageRank) but is assigned highest ranking score by our ranking algorithm. Page P-81 belongs to category 2 having three backward links and fourteen forward links. But since backward links of P-81 page are few and also all are not related to target topic so shows zero or very small ranking score in many LAR algorithms whereas our ranking algorithm equally considers all the three parameters (page content, backward links, forward links) for computing a page rank score and hence compute non-zero ranking score for P-81 since the content of P-81 is related to the target topic. Similar is the case with P-56, P-135 (belonging to category 2 having zero and one backward link respectively), P-74 (belonging to category 6 with neither page content nor backward link related to given topic). All are having zero or low ranking score in all LAR algorithms and non-zero or high ranking score in our ranking algorithm. P-119 is scored high in all LAR algorithms and low in our ranking algorithm since as compared to others web pages it's neither page content nor backward links are related to target topic. It has only two forward links referring to target topic and hence scored low by our algorithm. The ranking score results obtained by our ranking algorithm shows maximum similarity with the ranking scores obtained by other algorithms as shown in table 2.2 and 2.5 and also our algorithm shows better ranking results by assigning non-zero ranking values to these web pages.

The high dependency of ranking score of a web page on backward links is reduced in our ranking algorithm as shown by query results of "*Computational Complexity*" in table 3.2 of Appendix. Web Page P-55 is having high rank score in all LAR algorithms as compared to our ranking algorithm which assign low rank score to it. The reason is that P-55 belongs to category 7 having only backward links related to target topic. It has twenty forward links but neither forward links nor the page text content belongs to the target topic and hence is ranked low. P-55 is the home page of SDSC (SAN DIEGO Supercomputer Centre) which is not related to "Computational Complexity". Zero rank score is shown by all LAR algorithms including PageRank for Web pages P-31, P-33, P-28 (ref. table 3.2 and 3.5) while our ranking algorithm assigns small non-zero ranking score to them. The reason is that ranking scores computed by LAR algorithms depends on the nature of the backward links of web pages and these pages does not have any backward link and hence computes zero rank score, while our ranking algorithm equally considers all the three parameters (page content, backward links, forward links) for computing a page rank score and shows non-zero rank scores since the text content of pages P-31 and P-33 and the forward link of P-28 are related to topic "Computational Complexity" . Similarly page P-21 is ranked zero in many LAR algorithms due to zero forward link (poor hub) whereas our ranking algorithm computes non-zero ranking score since the page content and its backward links are related to the target topic.

VI. Conclusion

This paper describes a method for learning web structure to classify web documents and demonstrates the usefulness of considering the text content information of backward links and forward hyperlinks for page ranking. We also show that utilizing only extended anchor text from documents that link to the target document or while just considering the words and phrases on the target pages (full-text) does not yield very accurate results. In this paper, we analyze the similarity of the word usage at single level link distance from the page of interest and demonstrate that content of words in the linked pages enables more efficient indexing and searching. The new proposed method efficiently reduces the limitations of the some already existing Link Analysis algorithms while computing the rank of the retrieved web pages and the results obtained by the proposed method are not biased towards in-degree of the target page. Also the rank scores obtained shows non-zero values hence help to rank the web pages more accurately.

REFERENCES

[1] Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, & Panayiotis Tsaparas, "Finding Authorities and Hubs from link structures on the World Wide Web", in Proceedings of the 10th WWW Conference, Hong Kong, 2001, pp. 415-429.

[2] Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, & Panayiotis Tsaparas, "Link analysis ranking: algorithms, theory, and experiments", in ACM Trans. Inter. Tech., 5(1) , 2005, pp. 231-297.

[3] Ayman Farahat, Thomas Lofaro, Joel C. Miller, Gregory Rae, & Lesley A. Ward, "Authority Rankings from HITS, PageRank, and SALSA: Existence, Uniqueness, and Effect of Initialization", in SIAM J. Science Computing, 27(4), 2006, pp. 1181-1201.

[4] Boleslaw K. Szymanski, & Ming-shu Chung, "A method for Indexing Web Pages Using Web Bots", in Proceedings of the International Conference on Info-Tech Info-Net ICII'2001, Beijing, China, IEEE CS Press, 2001, pp. 1-6.

[5] Carriere J., & Kazman R., "Web query: Searching and visualizing the web through connectivity", in Proceedings of the 6th International World Wide Web conference, Santa Clara, California, 1997, pp. 1-14.

[6] David Gibson, Jon Kleinberg, & Prabhakar Raghavan, "Inferring Web Communities from Link Topology", in Proceedings of the 9th Conference on Hypertext and Hypermedia, 1998, pp. 225-234.

[7] D. Rafiei, & A. Mendelzon," What is this page known for? Computing web page reputations", in Proceedings of the 9th International World Wide Web Conference, Amsterdam, Netherlands, 2000, pp. 823–835.

[8] Jeromy Carriere, & Rick Razman," Webquery: Searching and visualizing the web through connectivity", in Proceedings of the 6th International WWW conference, Computer Networks and ISDN Systems,29, 1997, pp. 1257-1267.

[9] Jon M. Kleinberg, "Authoritative sources in a hyperlinked environment", in Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, 46(5), 1992, pp. 604-632.

[10] Justin Zobel, & Alistair Moffat, "Inverted Files for Text Search Engines", ACM Computing Surveys, 38 (2), 2006, pp. 1-56.

[11] Kiduk Yang, "Combining text-and link-based retrieval methods for Web IR", in Proc. of 10th Text REtrieval Conference, 2001, pp. 609—618.

[12] Longzhuang Li, Yi Shang, & Wei Zhang, "Improvement of HITS-based Algorithms on Web Documents", in Proceedings of the 11th international conference on World Wide Web, 2002, pp. 527-537.

[13] Massimo Marchiori, "The quest for correct information on the web: Hyper search engines", in Proceedings of the 6th International WWW Conference, 1997, pp. 265-274.

[14] Monika R. Henzinger, & Krishna Bharat, "Improved algorithms for topic distillation in a hyperlinked environment", in Proceedings of the 21st International ACM SIGIR conference on Research and Development in IR, 1998, pp. 104-111.

[15] Nadav Eiron, & Kevin S. McCurley, "Analysis of Anchor Text for Web Search", in Proc of the 26th annual international ACM SIGIR conference on Research and development in IR, 2003, pp. 459 – 460.

[16] Peter Pirolli, James Pitkow, & Ramana Rao, "Silk from a sow's ear: Extracting usuable structures from the web", in Proceedings of ACM SIGCHI conference on Human Factors in computing, 1996, pp. 118-125.

[17] Prem Chand Saxena, & Namita Gupta, "Quick Text Retrieval Algorithm Supporting Synonyms Based on Fuzzy Logic", in Computing Multimedia and Intelligent Techniques, 2(1), 2006, pp. 7-24.

[18] R. Lempel, & S. Moran, "The stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect", in Proc. of the 9th International World Wide Web Conference, Amsterdam, Netherlands, 2000, pp. 387-401.

[19] Sergey Brin, & Lawrence Page, "The anatomy of a large-scale hypertextual web search engine", in Proceedings of the 7th International WWW Conference, 30(1), 1998, pp. 107-117.

[20] Soumen Chakrabarati, Byron Dom, David Gibson, Jon M. Kleinberg, Prabhakar Raghavan, & Sridhar Rajagopalan, "Automatic resource list compilation by analyzing hyperlink structure and associated text", in Proc. of the 7th International WWW conference, 30(1-7), 1998, pp. 65-74.

[21] Thijs Westerveld, Wessel Kraaij, & Djoerd Hiemstra, "Retrieving Web Pages using Content, Links, URLs and Anchors", in Proceedings of the Tenth Text REtrieval Conference (TREC 2001), 2001, pp. 663-672.

[22] Yitong Wang , & Masaru Kitsuregawa, "Evaluating contents-link coupled web page clustering for web search results", in Proc. of the 11th Int. Conference on Web Information Systems Engineering (WISE'01, 2001.

[23] Zheng Chen, Shengping Liu, Liu Wenyin, Geguang Pu, & Wei-Ying Ma, "Building a web Thesaurus from web Link Structure", in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, 2003, pp. 48-55.

APPENDIX

1. COMPUTATIONAL GEOMETRY

Table 3.1. Showing detail information of ten web pages retrieved against query "Computational Geometry"

Web Page	Category	URL	TITLE
P-5	5	http://www.cs.duke.edu/CGC/workshop97.html	Second CGC Workshop on Computational Geometry
P-10	6	http://jeff.cs.mcgill.ca/cgm.html	Computational Geometry Lab at McGill
P-12	7	http://cs.smith.edu/~orourke/books/discrete.html	Handbook of Discrete and Computational Geometry
P-20	5	http://www.ics.uci.edu/~eppstein/266	Computational Geometry
P-21	5	http://dimacs.rutgers.edu/Volumes/Vol06.html	Volume 6 "Discrete and Computational Geometry: Papers from the DIMACS Special Year", Goodman, Pollack & Steiger, Eds.
P-23	5	http://archives.math.utk.edu/topics/computationalGeom.html	Mathematics Archives - Topics in Mathematics - Computational Geometry
P-27	3	http://www.siam.org/meetings/archives/an97/ms8.htm	MS8 Computational Geometry Approaches to Mesh Generation
P-28	5	http://www.math-inst.hu/staff/geometry.html	Convex and Computational Geometry research group
P-31	5	http://www.sonic.net/~sjl/compgeom.html	Computational Geometry
P-50	5	http://www.risc.uni-linz.ac.at/projects/basic/cgal	Computational Geometry Algorithms Library

Table 3.2. Showing the Rank Scores of ten web pages obtained by different LAR algorithms

Page No.	Kleinberg	pSALSA	SALSA	HubAvg	AThresh	HThresh	FThresh	BFS	New Method
P-5	0.001454	0.003289	0.003055	0.000082	0.011785	0.073141	0.001335	243.070312	14.837037
P-10	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	5.000000
P-12	0.001450	0.001771	0.001645	0.000023	0.011436	0.073141	0.001246	241.453125	7.800000
P-20	0.003105	0.001265	0.001175	0.000059	0.028738	0.127205	0.002447	250.835938	15.500000
P-21	0.000216	0.000506	0.000470	0.000004	0.002680	0.011887	0.000267	138.656250	5.500000

P-23	0.003365	0.002530	0.002350	0.000138	0.035451	0.188979	0.003601	245.226562	16.416667
P-27	0.000054	0.000506	0.000470	0.000002	0.000760	0.002674	0.000067	92.156250	6.000000
P-28	0.002133	0.001265	0.001175	0.000044	0.023483	0.119727	0.002345	201.992188	4.000000
P-31	0.003403	0.001771	0.001645	0.000097	0.035314	0.188979	0.003329	243.351562	18.839286
P-50	0.000423	0.000506	0.000470	0.000048	0.007112	0.022403	0.000693	150.828125	7.000000

Table 3.3. List the web pages in decreasing order of their Rank Scores ** ** Italics and underline pages shows web pages with zero rank score

Kleinberg	pSALSA	SALSA	HubAvg	AThresh	HThresh	FThresh	BFS	New Method
P-31	**P-5**	**P-5**	**P-23**	**P-23**	**P-31**	**P-23**	**P-20**	**P-31**
P-23	**P-23**	**P-23**	**P-31**	**P-31**	**P-23**	**P-31**	**P-23**	**P-23**
P-20	**P-31**	**P-31**	**P-5**	**P-20**	**P-20**	**P-20**	**P-31**	**P-20**
P-28	**P-12**	**P-12**	**P-20**	**P-28**	**P-28**	**P-28**	**P-5**	**P-5**
P-5	**P-28**	**P-28**	**P-50**	**P-5**	**P-5**	**P-5**	**P-12**	**P-12**
P-12	**P-20**	**P-20**	**P-28**	**P-12**	**P-12**	**P-12**	**P-28**	**P-50**
P-50	**P-27**	**P-27**	**P-12**	**P-50**	**P-50**	**P-50**	**P-50**	**P-27**
P-21	**P-50**	**P-50**	**P-21**	**P-21**	**P-21**	**P-21**	**P-21**	**P-21**
P-27	**P-21**	**P-21**	**P-27**	**P-27**	**P-27**	**P-27**	**P-27**	**P-10**
P-10	*P-10*	*P-10*	*P-10*	*P-10*	*P-10*	*P-10*	*P-10*	P-28

Table 3.4. Showing the relative ranking of web pages in different LAR algorithms ** Gray boxes shows web pages with zero score

Page No.	Kleinberg	pSALSA	SALSA	HubAvg	AThresh	HThresh	FThresh	BFS	New Method
P-5	9	1	1	6	6	9	6	4	9
P-10	6	6	6	9	9	6	9	6	6
P-12	4	9	9	1	4	4	4	9	4
P-20	8	3	3	4	8	8	8	1	1
P-21	1	8	8	10	1	1	1	3	3
P-23	3	4	4	8	3	3	3	8	10
P-27	10	7	7	3	10	10	10	10	7
P-28	5	10	10	5	5	5	5	5	5
P-31	7	5	5	7	7	7	7	7	2
P-50	2	2	2	2	2	2	2	2	8

Table 3.5. Showing the Rank Scores of ten web pages obtained by different Ranking algorithms

Page No.	PageRank	Alexa Rank	AltaVista Results	AllTheWeb Results
P-5	5	5528	108	105
P-10	5	9232	42	42
P-12	5	0	59	53
P-20	4	245091	65	65
P-21	0	4555	55	55
P-23	5	10568	88	84
P-27	0	220804	37	37
P-28	5	1801971	46	46
P-31	0	8069605	50	50
P-50	4	71944	58	52

2. ABORTION

Table 3.6. Showing detail information of ten web pages retrieved against query "Abortion"

Web Page	Category	URL	TITLE
P-12	2	http://www.feminist.org/rrights/medical.html	Reproductive Rights - Medical Abortion
P-14	1	http://www.wcla.org	Welcome to the Westchester Coalition for Legal Abortion
P-56	2	http://www.govspot.com/issues/abortion.htm	Abortion: GovSpot Issues
P-60	1	http://www.rightgrrl.com/carolyn/peasoup.html	Rightgrrl - Abortion and Pea Soup - Post Abortion Syndrome
P-74	6	http://www.rickross.com/groups/abortion.html	Rick Ross: Anti-Abortion Extremists
P-79	2	http://enquirer.com/columns/crowley/1999/04/25/pcr_abortion_foe_eases.html	Abortion foe eases up a bit with reporters
P-81	2	http://www.religioustolerance.org/abo_viol.htm	Violence at US Abortion Clinics
P-88	3	http://www.hopemedical.com/5.htm	Hope Medical Group for Women, Shreveport Louisiana
P-119	6	http://www.crusadeforlife.org	Crusade for Life, Christians Protecting Families from Abortion and Euthanasia
P-135	2	http://www.onlineathens.com/1998/100398/1003.a3pill.html	Moving to speed spending bills, GOP drops ban on abortion pill

Table 3.7. Showing the Rank Scores of ten web pages obtained by different LAR algorithms

Page No.	Kleinberg	pSALSA	SALSA	HubAvg	AThresh	HThresh	FThresh	BFS	New Method
P-12	0.000000	0.000104	0.000099	0.000000	0.000001	0.002997	0.000001	154.546875	4.000000
P-14	0.000009	0.003111	0.002966	0.000002	0.000063	0.169263	0.000051	389.625000	11.705882

P-56	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	3.555556
P-60	0.000000	0.000311	0.000297	0.000000	0.000000	0.001626	0.000000	159.046875	3.000000
P-74	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	13.500000
P-79	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	2.000000
P-81	0.000000	0.000311	0.000297	0.000000	0.000000	0.001261	0.000000	139.089844	17.875000
P-88	0.000003	0.000104	0.000099	0.000001	0.000020	0.017972	0.000011	261.003906	2.000000
P-119	0.000000	0.000311	0.000297	0.000000	0.000001	0.011122	0.000001	210.937500	1.000000
P-135	0.000000	0.000104	0.000099	0.000000	0.000000	0.000002	0.000000	42.434082	7.000000

Table 3.8. List the web pages in decreasing order of their Rank Scores ** Italics and underline pages shows web pages with zero rank score

Kleinberg	pSALSA	SALSA	HubAvg	AThresh	HThresh	FThresh	BFS	New Method
P-14	**P-14**	**P-14**	**P-14**	**P-14**	**P-14**	**P-14**	**P-14**	**P-81**
P-88	**P-119**	**P-119**	**P-88**	**P-88**	**P-88**	**P-88**	**P-88**	**P-74**
P-135	**P-81**	**P-81**	***P-135***	**P-119**	**P-119**	**P-119**	**P-119**	**P-14**
P-81	**P-60**	**P-60**	***P-81***	**P-12**	**P-12**	**P-12**	**P-60**	**P-135**
P-119	**P-88**	**P-88**	***P-119***	***P-56***	**P-60**	***P-56***	**P-12**	**P-12**
P-79	**P-135**	**P-135**	***P-79***	***P-135***	**P-81**	***P-135***	**P-81**	**P-56**
P-56	**P-12**	**P-12**	***P-56***	***P-60***	**P-135**	***P-60***	**P-135**	**P-60**
P-12	***P-79***	***P-79***	***P-12***	***P-74***	***P-74***	***P-74***	***P-74***	**P-88**
P-74	***P-56***	***P-56***	***P-74***	***P-79***	***P-79***	***P-79***	***P-79***	**P-79**
P-60	***P-74***	***P-74***	***P-60***	***P-81***	***P-56***	***P-81***	***P-56***	**P-119**

Table 3.9. Showing the relative ranking of web pages in different LAR algorithms ** Gray boxes shows web pages with zero score

Page No.	Kleinberg	pSALSA	SALSA	HubAvg	AThresh	HThresh	FThresh	BFS	New Method
P-12	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**7**
P-14	**8**	**9**	**9**	**8**	**8**	**8**	**8**	**8**	**5**
P-56	**10**	**7**	**7**	**10**	**9**	**9**	**9**	**9**	**2**
P-60	**7**	**4**	**4**	**7**	**1**	**1**	**1**	**4**	**10**
P-74	**9**	**8**	**8**	**9**	**3**	**4**	**3**	**1**	**1**
P-79	**6**	**10**	**10**	**6**	**10**	**7**	**10**	**7**	**3**
P-81	**3**	**1**	**1**	**3**	**4**	**10**	**4**	**10**	**4**
P-88	**1**	**6**	**6**	**1**	**5**	**5**	**5**	**5**	**8**
P-119	**5**	**3**	**3**	**5**	**6**	**6**	**6**	**6**	**6**
P-135	**4**	**5**	**5**	**4**	**7**	**3**	**7**	**3**	**9**

Table 3.10. Showing the Rank Scores of ten web pages obtained by different Ranking algorithms

Page No.	Page Rank	Alexa Rank	AltaVista Results	AllTheWeb Results
P-12	5	289302	157	151
P-14	4	0	178	175
P-56	4	237310	26	26
P-60	2	3029642	26	26
P-74	4	54009	114	107
P-79	0	59307	30	30
P-81	2	21482	892	819
P-88	3	16617134	30	30
P-119	3	5773513	149	148
P-135	0	41030	4	4

3. COMPUTATIONAL COMPLEXITY

Table 3.1. Showing detail information of ten web pages retrieved against query “Computational Complexity”

Web Page	Category	URL	TITLE
P-2	6	http://www-math.uni-paderborn.de/~aggathen/cc	cc homepage
P-14	6	http://www.cse.buffalo.edu/pub/WWW/faculty/regan/ccc98	1998 IEEE Conference on Computational Complexity
P-18	6	http://www.cs.utep.edu/longpre/complexity.html	IEEE Conference on Computational Complexity
P-21	3	http://www.cs.rochester.edu/courses/descriptions/286.html	CSC 286/486: Computational Complexity
P-28	6	http://gort.ucsd.edu/newjour/e/msg02611.html	Electronic Colloquium on Computational Complexity
P-31	2	http://www.informatik.uni-hamburg.de/TGI/pnbib/m/magott_j4.html	Computational Complexity of Algorithms and Problems of Minimal Cycle Time for Chosen Classes of Petri Nets.
P-33	6	http://www.cs.princeton.edu/courses/archive/spr00/cs522/assignments.html	CS 522:Computational Complexity
P-38	6	http://elib.cs.sfu.ca/cs-journals/P-Birkhauser/J-Birkhauser-CC.html	Computational Complexity
P-43	5	http://synapse.cs.byu.edu/~dan/complex.html	Computational Complexity
P-55	7	http://www.sdsc.edu	SDSC: A National Laboratory for Computational Science and Engineering

Table 3.2. Showing the Rank Scores of ten web pages obtained by different LAR algorithms

Page No.	Kleinberg	pSALSA	SALSA	HubAvg	AThresh	HThresh	FThresh	BFS	New Method
P-2	0.000000	0.001238	0.001845	0.000000	0.000000	0.000000	0.000000	2.000000	18.583333
P-14	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	26.000000
P-18	0.000313	0.005569	0.005416	0.000772	0.030666	0.006308	0.004812	100.720703	11.200000
P-21	0.000000	0.001238	0.001845	0.000000	0.000000	0.000000	0.000000	2.000000	7.000000
P-28	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	5.000000
P-31	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.250000
P-33	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	4.500000
P-38	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	6.000000
P-43	0.000313	0.001238	0.001204	0.000791	0.009146	0.001954	0.001236	67.013672	5.866667
P-55	0.000022	0.005569	0.005416	0.000057	0.002007	0.000654	0.000365	52.812500	4.769737

Table 3.3. List the web pages in decreasing order of their Rank Scores ** Italics and underline pages shows web pages with zero rank score

Kleinberg	pSALSA	SALSA	HubAvg	AThresh	HThresh	FThresh	BFS	New Method
P-18	**P-55**	**P-55**	**P-43**	**P-18**	**P-18**	**P-18**	**P-18**	**P-14**
P-43	**P-18**	**P-18**	**P-18**	**P-43**	**P-43**	**P-43**	**P-43**	**P-2**
P-55	**P-2**	**P-2**	**P-55**	**P-55**	**P-55**	**P-55**	**P-55**	**P-18**
P-38	**P-43**	**P-21**	***P-38***	***P-38***	***P-38***	***P-38***	**P-2**	**P-21**
P-14	**P-21**	**P-43**	***P-14***	***P-14***	***P-14***	***P-14***	**P-21**	**P-38**
P-2	***P-28***	***P-28***	***P-2***	***P-2***	***P-2***	***P-2***	***P-14***	**P-43**
P-21	***P-14***	***P-14***	***P-21***	***P-21***	***P-21***	***P-21***	***P-28***	**P-28**
P-28	***P-31***	***P-31***	***P-28***	***P-28***	***P-28***	***P-28***	***P-31***	**P-55**
P-31	***P-33***	***P-33***	***P-31***	***P-31***	***P-31***	***P-31***	***P-33***	**P-33**
P-33	***P-38***	***P-38***	***P-33***	***P-33***	***P-33***	***P-33***	***P-38***	**P-31**

Table 3.4. Showing the relative ranking of web pages in different LAR algorithms ** Gray boxes shows web pages with zero score

Page No.	Kleinberg	pSALSA	SALSA	HubAvg	AThresh	HThresh	FThresh	BFS	New Method
P-2	3	10	10	9	3	3	3	3	2
P-14	9	3	3	3	9	9	9	9	1
P-18	10	1	1	10	10	10	10	10	3
P-21	8	9	4	8	8	8	8	1	4
P-28	2	4	9	2	2	2	2	4	8
P-31	1	5	5	1	1	1	1	2	9
P-33	4	2	2	4	4	4	4	5	5
P-38	5	6	6	5	5	5	5	6	10
P-43	6	7	7	6	6	6	6	7	7
P-55	7	8	8	7	7	7	7	8	6

Table 3.5. Showing the Rank Scores of ten web pages obtained by different Ranking algorithms

Page No.	PageRank	Alexa Rank	AltaVista Results	AllTheWeb Results
P-2	5	0	56	56
P-14	0	10331	39	39
P-18	4	50041	83	77
P-21	0	13255	51	51
P-28	0	5425	41	41
P-31	0	8907	42	42
P-33	0	5014	18	18
P-38	0	11743	42	42
P-43	0	21398378	44	45
P-55	8	125692	35900	34200

Prof. J.P. Gupta obtained Master's Degree in Electronics & Communication Engineering with the Gold Medal in 1973 from the University of Roorkee, India. He obtained his Doctorate Degree in Computer Engineering from the University of Westminster, London under the Commonwealth Scholarship Award. He held a position of Professor serving the University of Roorkee (now IIT, Roorkee) for over 25 years. Prof. Gupta was the Member Secretary, All India Council for Technical Education (AICTE) (1994-98). Presently he is working as Vice-Chancellor of JIIT since July 2005. Distinguished in Computer Engineering, Prof. Gupta has a vast research experience, besides being involved in numerous Consultancy and Research & Development activities.

Prem Chand Saxena earned his Master of Science Degree from Delhi University, Delhi, India in 1968. Then he received his Doctor of Philosophy Degree in Operational Research from Delhi University, Delhi, India in 1974. Now he is working as Professor, Computer Science at School of Computer & Systems Sciences, Jawaharlal Nehru University, Delhi, India.
He has supervised 13 PhD students and guided 85 M. Tech. Dissertations. His areas of research and interest are Database Management System, Data Communication, Distributed Systems, Data Mining, Mobile Computing, Networking and Multimedia.

Namita Gupta earned her Master in Computer Applications from Maharashi Dayanand University, Rohtak, Haryana, India in 1998. Now she is working as Assistant Professor, Computer Science, Maharaja Agarsen Institute of Technology, GGSIP University, Delhi, India. Her topics of interest are Data Mining, Software Engineering, DBMS, Operating System. Currently she is doing research work in Text mining.