HierarchicalRank: Webpage Rank Improvement Using HTML TagLevel Similarity

Dilip Sharma and Deepak Ganeshiya
Department of Computer Engineering and Applications, GLA University Mathura, India

Abstract: *In the past researches, two types of algorithms are introduced that are query dependent and query independent, works online or offline. PageRank Algorithm works offline independent to query while Hyperlink-Induced Topic Search (HITS) algorithm woks online dependent on query. One of the problems of these algorithms is that, division of the rank is based on number of inlinks, outlinks and different parameters used in hyperlink analysis which is dependent or independent to webpage content with the problem of topic drift. Previous researches were focused to solve this problem using the popularity of the outlink webpages. In this paper a novel algorithm for popularity measure is proposed based on similarity between query and Hierarchical text extracted from source and target webpage using Hyper Text Markup Language (HTML) tags importance parameter. In this paper, result of proposed method is compared with PageRank Algorithm and Topic Distillation with Query Dependent Link Connections and Page Characteristics results.*

Keywords: *Web mining, web graph, hyperlink analysis, connectivity, pagerank, HTML tags.*

Received July 21, 2014; accepted October 14, 2014

1. Introduction

With the growth of technology in the field of internet, number of web pages on World Wide Web (WWW) is increasing day by day. So, size of the information on WWW grows very rapidly. Thus the necessity of searching useful and required information comes into front. For this purpose, it is a need to mine required information from the internet, Web mining is used to retrieve the relative information from the internet. For this purpose, search engines are used that works as interface between user and WWW. Different search engines provide required information to the user based on different ranking algorithms. In the WWW large number of webpages are linked to each other directly or indirectly through hyperlinks and create a web graph [2, 7, 13, 18]. For retrieving information from the web three classes of Web mining are used that are:

1) Web Content Mining (WCM).
2) Web Structure Mining (WSM).
3) Web Usage Mining (WUM) [5].

WCM related to extracting the required content from the webpage. WSM is related to relationships among webpages and using linking information for webpage ranking through hyperlink analysis. WUM is used to extract the user's profile (for example number of clicks for a particular topic). These three mining techniques have been used by various researchers in their work such as WSM used in PageRank Algorithm [1] and Weighted PageRank Algorithm [20]. Combination of these techniques is used in various researches related to improvement in ranking. WSM is used with WUM in 'PageRanking based on number of visit of link of webpage' by Kumar *et al*. [10], 'Weighted PageRanking based visit of links' by Tyagi and Sharma [17] and 'An Improved Page Rank Algorithm Based on Optimized Normalization Technique' by Dubey and Roy [4]. Figure 1 shows the classification of the various research works based on different web mining techniques.

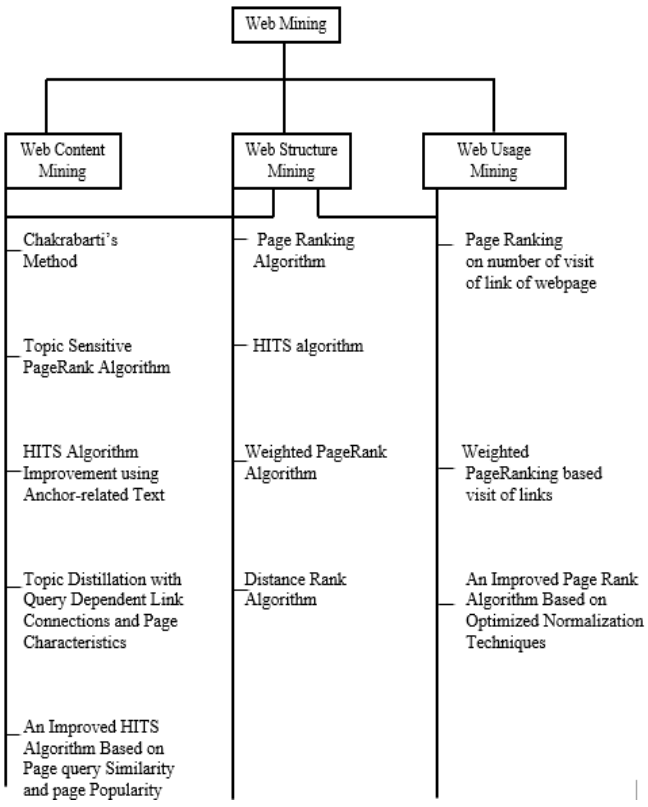

Figure 1. Classification of webpage ranking algorithms based on web mining techniques.

In the previous research, hyperlink analysis is the most used approach which is used to find out various resources from webpages around hyperlink for webpage ranking [7]. Topic distillation approach is used to find key resources from the webpage to extract the topic of that webpage. These two approaches are used in various researches in the area of ranking. In the source webpage; anchor text is the most important resource which is used for topic distillation [19]. Anchor text and its related texts are used in the various researches for improvement of relevancy of the result with respect to the given query [15]. This text is used to measure popularity of the target webpage and thereafter this popularity is used for measuring the rank value of target webpage. Previous research related to ranking algorithms, divides the ranking algorithms into two classes: First, *Query independent algorithm* that works offline independent to user's query, for example PageRank Algorithm [1]. Second, *Query dependent algorithm* that works online dependent to user's query, for example Hyperlink-Induced Topic Search (HITS) Algorithm [8, 9]. There are some researches which uses different types of anchor related text sets are given by Chakrabarti *et al.* [3], Zhang *et al.* [21], and Tao and Zuo [16]. In this paper Hierarchical text and Hyper Text Markup Language (HTML) tag importance parameter is proposed which is used for measurement of popularity of target webpage. Next section discusses the related work in the field of webpage ranking and topic distillation which uses WCM and WSM to calculate the webpage importance. Experimental methodology is discussed in section 4. And the experiment results are discussed in section 5. And finally the conclusions and future work are discussed in section 7.

2. Related Work

In the field of webpage ranking, different algorithms were introduced by various researchers in their research using different classes of web mining. Some of them are discussed below:

Page Ranking Algorithm [1] is the traditional algorithm to provide indexing to webpage ranking, which is given by S. Brin and L. Page. The rank of a webpage is depending on the rank of its backlink webpages. The strength of this algorithm is that rank of a webpage is done on the basis of importance of backlink webpages. Its limitation is that ranking is done offline at the indexing time.

HITS algorithm [9] proposed method which ranks the webpages query dependent and works online. It computes ranking based on two types of scores that are hub and authority score. Hub webpages have links to the authority webpages which contains useful and required information. It's strength is that it works online, and limitation is the problem of topic drift, which refers that returned webpage may not be relevant to the query.

Chakrabarti's *et al*. method [3] improves HITS algorithm by introducing 50 words window text around the anchor text for measuring the popularity of target webpage. In this algorithm Query weight is calculated based on the occurrences of query in the 50 words window text. In this method webpage ranking is based on query weight with hub and authority scores of HITS algorithm [9]. Strength of this algorithm is easy computation of query_wt and limitation is that it is not necessary that 50 words window text is related to the target page.

Topic Sensitive PageRank Algorithm [6], is the next version of PageRank Algorithm based on multiple rank vectors, it computes a set of rank vectors offline with respect to 16 topics selected from Open Directory Project (ODP). There is a problem in PageRank algorithm that webpages which have more inlinks are bound to get assigned more rank value for queries for which they have no related information which is avoided by this method. Similarity between query and the 16 topics are calculated at query time. Strength of this algorithm is thatit avoids the problem of heavily linked pages getting highly ranked and limitation is that it satisfies only local (16 topics) authority not global authority.

Query-sensitive self-adaptable ranking algorithm [16], introduced rank value which satisfies global as well as local authority with respect to query. In this algorithm query sensitiveness approach is used for satisfying local importance, using a voting procedure which uses result list of webpages to build a voting set V_{Set}, in this set all webpages have same right to vote. A webpage $doc_{is,}$ can vote to webpage $doc_{it.}$ if they satisfy the following conditions.

- If there is a link from doc_{is} to doc_{it}
- If doc_{it}'s title, is present in the content of doc_{is}
- If doc_{is}'s information is referred from the content of doc_{it}.

Only one vote is counted if at least two conditions from above mentioned three conditions satisfies simultaneously. The calculation of webpage ranking score is based on query sensitiveness value of a webpage and their respective PageRank value.

Weighted PageRank (WPR) Algorithm [20] is the improved version of PageRank algorithm by introducing the weighted scheme to the link based on the popularity of the target webpage which is defined on the basis of inlink and outlink weighting scheme. It's weakness is that it ranks the webpages on the basis of popularity which may be distinguished easily as its page rank is directly proportional to the number of its inlinks.

Distance Rank Algorithm [12] is based on reinforcement learning and depends on distance between webpages which repeatedly calculates

distance based on previous distance value. Distance between webpages is defined as punishment which calculates based on logarithm of number of the outlinks of the inlink webpages.

Weakness of Kleinberg's HITS algorithm is topic drift problem. To resolve this weakness the Improvement in HITS algorithm is achieved using anchor related text [15]. The anchor related text contains two types of texts, Upper Level Text (ULT) and Lower Level Text (LLT). LLT extracted using three types of tags: parent tags, sibling tags and relative tags with respect to position of the anchor tag. ULT exists in the <title>, <table> tag and header tags. This anchor related text is used to calculate the importance of the webpage by using Chakrabarti's *et al*. Method [3].

Topic distillation approach is defined as finding the resources from webpages that defines the topic of the webpage. This approach is broadly categorized into two classes: evidence based on link structure and evidence relates to webpage characteristics. In this work the initial search results are reordered based on a

Combination of external evidences and the original similarity score of a webpage. In this approach weighted linear combination method is used to combine the external evidences and the original similarity score. And finally the initial result list is reordered based on modified similarity scores of webpages. Strength of this algorithm is that different evidences provide the gain in web page rank and limitation is that the Anchor density provides less improvement in web page rank.

Improved HITS algorithm [11] uses content from source and target webpage torank the target webpage.In this approach the rank calculations based on hub and authority scores and the similarity between source webpage and target webpage. The popularity score of the target webpage is calculated using weighted scheme of Weighted PageRank Algorithm [20]. Strength of this algorithm is that similarity measured is based on similarity between content of source and target webpage. Limitation is that complete source page content is not useful for similarity calculation.

3. Problem Formulation

Above section of this paper discusses the various researches in the area of ranking, with introducing different web mining techniques. The main issues in webpage ranking are division of source webpage rank to target and measuring popularity of target webpage is also discussed in previous section. In the previous researches the popularity of the target webpage is measured with respect to different types of text sets extracted from the source and target webpage. In the next section, the popularity of target webpage is measured by using the similarity between query and Hierarchical text using HTML tag importance parameter.

4. Experimental Methodology

Previously researchers have tried to improve the webpage ranking by interpolating different approaches like hyperlink analysis and topic distillation with different mining techniques which are discussed in second section. In introducing WCM, content of source and target webpage is used for measuring the popularity of target webpage and later this popularity is used in the calculation of target webpage rank. Here a methodology is proposed for extracting Hierarchical text using HTML tag importance parameter for measuring popularity of target webpage. Figure 2 shows the framework for proposed algorithm.

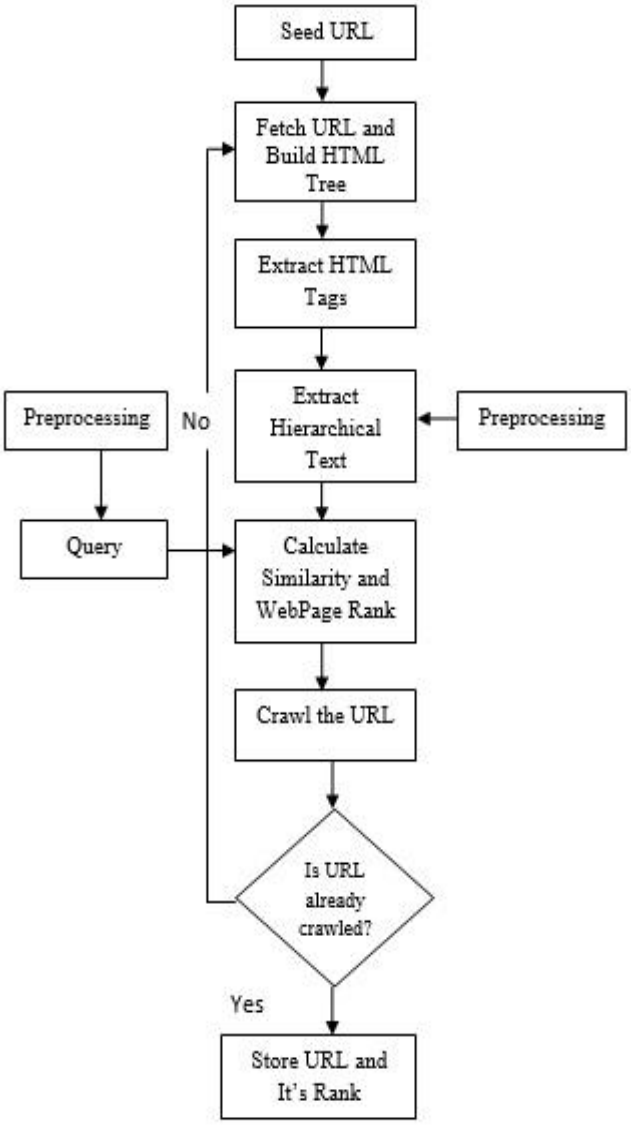

Figure 2. Framework for proposed hierarchical rank algorithm.

4.1. Result Measures and Base Lines

In terms of measuring the effectiveness of the result the precision of result list is used which is the ratio of number of relevant webpages that have been retrieved to the total number of webpages that the system has retrieved [14]. Three valuations are used to evaluate the performance of the system, these are mean

precision at each retrieved documents, mean R-precision, the precision after R documents are retrieved for particular query, and Mean Average Precision (MAP). Average precision for a single query is calculated by averaging the precision scores at each point in theresultlist where a relevant document is found.

4.2. Hierarchical Rank Algorithm

Here the algorithm is defined by following points with pre-processing the query and extracted webpage content for calculation of webpage ranking by applying following subroutine:

Algorithm 1: Preprocess(WebGraph G, Query Q)

```
String[] HierarchicalText = new String[4];
For each WebPage p in G do {
    Remove self-pointing hyperlinks;
    HierarchicalText[]=ExtractHierarchicalText(p);
    Remove stopwords from HierarchicalText[] and Q;
    Apply Stemming using Porter Stemmer on
    HierarchicalText[] and Q;
    Extract Kywords from HierarchicalText[] and Q;
    Find Similarity Measure using Q and
    HierarchicalText[]; }
Compute and store the HierarchicalRank of each
webpage p in G;
```

4.3. Extraction of Hierarchical Text

In this section, firstly investigate the Hierarchical text from the source and target webpage then extract the Hierarchical text. Lastly, the similarity measure between query and Hierarchical text is taken into account.

In the investigation of the Hierarchical text from the source webpage, primary HTML tags are used. These primary tags are parent tags which containsdifferent child tags as secondary tags such as <table>tag used as primary tag with<th>, <tr> and <td>as its child tags. Beside this other text from source as well as target webpage is also used with this Hierarchical text. Following subroutines are used for extraction of Hierarchical text.

Algorithm 2: ExtractHierarchicalText(WebPage q)

```
String[] HierarchicalText = new String[5];
For each BackLinkWebPage p of q do {
    Parse the WebPage p and find the tree level l which
    contains hyperlink p to q;
    HierarchicalText[0] = ExtractAnchorText(p, q);
    HierarchicalText[1] = ExtractTreeLevelText(p, l);
    HierarchicalText[2] = ExtractURLText(q);
    HierarchicalText[3] = ExtractTitleText(q);
    HierarchicalText[4] = ExtractOtherText(q);
Return HierarchicalText[5];
```

Algorihtm 3: ExtractAnchorText(WebPage p, WebPage q)

```
String AnchorText = NULL;
AnchorText = Extract text associated with anchor tag
<a> of hyperlink p to q;
Return AnchorText;
```

Algorihtm 4: ExtractTreeLevelText(WebPage p, TreeLevel l)

```
String TreeLevelText = NULL;
TreeLevelText = Extract text associated with primary
HTML tags present at l in p;
Return TreeLevelText;
```

Algorihtm 5: ExtractURLText(WebPage q)

```
String Url = NULL;
Url = Extract URL of q;
Return Url;
```

Algorihtm 6: ExtractTitleText(WebPage q)

```
String TitleText = NULL;
TitleText = Extract text associated with <title> HTML
tag of q;
Return TitleText;
```

Algorihtm 7: ExtractOtherText(WebPage q)

```
String OtherText = NULL;
OtherText = Extract other text from q associated with
different HTML tags except <title>;
Return OtherText;
```

4.4. Similarity Measure

In the previous subsection 4.3, Hierarchical text is obtained, thereafterfor a webpage HierarchicalRank is calculated using HTML tag importance parameter α. Values of this importance parameter with their respected HTML tags and associated different types of texts (which are obtained in Hierarchical text extraction discussed in section 4.3) are shown in Table 1. Importance parametervalues are categorized for different HTML tags based on the probability of occurrences of query in their respected associated texts. These importanceparametervalues are used to maximize the impact of occurrence of a query in text whose corresponding HTML tag has highest probability of finding the query. For a source webpage v and target webpage u, similarity measure is calculated by using following Equation:

$$S_{(v,u)} = \frac{\sum_{i=1}^{2} \alpha_i S_v^i + \sum_{i=3}^{5} \alpha_i S_u^i}{\sum_{i=1}^{2} S_v^i + \sum_{i=3}^{5} S_u^i} \quad (1)$$

Where $S_{(v,u)}$ is the similarity measure value which is used in computation of HierarchicalRank HR(u) for webpage u. Here i is used for five different texts which are extracted with extraction of Hierarchical text. S_v^i is the cosine similarity between query and corresponding text with respect to i and respected webpage u or v, and α_i is the HTML tag importance value of corresponding i^{th} text which is described in Table 1. For a webpage u HierarchicalRank HR(u) is calculated as following:

$$HR(u) = (1-d) + d_{v\varepsilon B(u)} \frac{HR(v)}{N_v} + S_{(v,u)} \qquad (2)$$

Here, *d* represents the dampening factor which is used in PageRank Algorithm [1] and *B*(*u*) is the set of backlink webpages of webpage *u*.

Table 1. Description of texts extracted in extraction of hierarchical text with their corresponding HTML tags.

Value of i	Type of Text	HTML tag and text Description	Tag Importance parameter (α) value
1	Anchor text	Text associated with anchor tag <a>	0.20
2	Hierarchical text	Text associated with Primary tags present at level of anchor tag <a> excluding anchor text	0.15
3	URL	URL text of target webpage	0.25
4	Title text	Text associated with <title> tag of target webpage	0.30
5	Other text	Text associated with all HTML tags of target webpage except <title> tag	0.10

5. Experimental Analysis

In the previous section the experimental method is discussed. For measuring the effectiveness of Hierarchical Rank algorithm, respective result is compared with other two algorithms PageRank Algorithm [1], Topic Distillation with Query Dependent Link Connections and Page Characteristics [19] results.

5.1. Data Set

For the analysis the result, the website of University of California at Berkeley's Web portalis used as the dataset which contains approximately five million web pages [12]. Along with this 10 queries are used that are: news and events, academic, student, course, scholarship, result, admission, placement, education and department. For an input query "news and events"

Result lists of relevant webpages carried out using PageRank Algorithm [1], Topic Distillation with Query Dependent Link Connections and Page Characteristics [19] and proposed Hierarchical Rank algorithm. In the result list relevancy order of the webpages precision values of top 10 webpages are obtained which are shown in following 3 Tables:

Table 2. List of webpages based on pagerank algorithm.

Rank	Webpage URL	Precision (%)
1	http://www.berkeley.edu/	100
2	http://newscenter.berkeley.edu	100
3	http://bconnected.berkeley.edu/	66.67
4	http://calparents.berkeley.edu/	50
5	http://alumni.berkeley.edu/california-magazine/just-in/2014-07-25/bones-pick-uc-berkeley-paleontologist-entices-diverse	40
6	http://alumni-friends.berkeley.edu/	33.33
7	http://directory.berkeley.edu	28.57
8	https://bspace.berkeley.edu/portal	25
9	http://diversity.berkeley.edu/	22.22
10	http://events.berkeley.edu/	30

Table 3. List of webpages based on topical distillation based on link evidence and webpage characteristics using anchor text, α_0=.2 for PR, α_1=.55 for Link Evidence, α_2=.25 for page feature.

Rank	Webpage URL	Precision (%)
1	http://newscenter.berkeley.edu/2014/07/29/university-librarian-reflects-on-a-transformative-era/	100
2	http://alumni.berkeley.edu/california-magazine/just-in/2014-07-25/bones-pick-uc-berkeley-paleontologist-entices-diverse	50
3	http://newscenter.berkeley.edu/2014/07/29/vision-correcting-displays/	66.67
4	http://newscenter.berkeley.edu/2014/07/29/happy-99th-charles-townes/	75
5	http://www.berkeley.edu/	80
6	http://blogs.berkeley.edu	66.67
7	http://blogs.berkeley.edu/2014/07/29/is-the-u-s-falling-behind-mexico-news-from-ambos-nogales/	57.14
8	http://newscenter.berkeley.edu	62.5
9	http://events.berkeley.edu/	66.67
10	http://bconnected.berkeley.edu/	60

Table 4. List of webpages based on hierarchical rank algorithm.

Rank	Webpage URL	Precision (%)
1	http://newscenter.berkeley.edu/2014/07/29/university-librarian-reflects-on-a-transformative-era/	100
2	http://newscenter.berkeley.edu/2014/07/29/vision-correcting-displays/	100
3	http://newscenter.berkeley.edu/2014/07/29/happy-99th-charles-townes/	100
4	http://alumni.berkeley.edu/california-magazine/just-in/2014-07-25/bones-pick-uc-berkeley-paleontologist-entices-diverse	75
5	http://www.berkeley.edu/	80
6	http://newscenter.berkeley.edu	83.33
7	http://blogs.berkeley.edu/2014/07/29/is-the-u-s-falling-behind-mexico-news-from-ambos-nogales/	71.42
8	http://events.berkeley.edu/	75
9	http://bconnected.berkeley.edu/	66.67
10	http://newscenter.berkeley.edu/2014/07/24/wildlife-decline-drives-social-conflict/	70

In the analysis of the proposed approach, the precision values are obtained. With the analysis of these values Figure 3 shows the precision curve of result lists at each retrieved URL using different algorithms and Table 5 presents the Mean Average Precision (MAP) values obtained with respect to these algorithms.

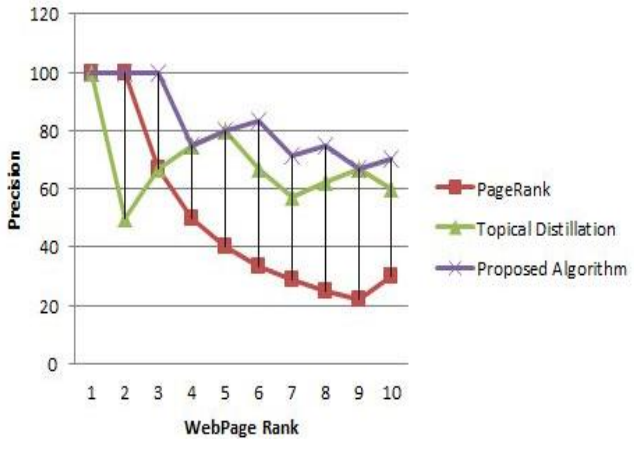

Figure 3. The precision curve with respect to the result obtained from different algorithms.

Table 5. Mean average precision (MAP) values with respect to different algorithms.

Algorithm	MAP (%)
PageRank Algorithm	45.07
Topical Distillation	62.24
Hierarchical Rank Algorithm	74.67

In the proposed algorithm different types of texts are extracted in the extraction of Hierarchical text (discussed in section 4.3). Table 6 and Figure 4 shows the MAP and gain (percentage of increment of MAP) of the result list obtained with respect to proposed algorithm using different texts and their corresponding HTML tag importance parameter α (described in table 1).

Table 6. Mean Average Precision (MAP) and their gain values with respect to different texts.

Text set	Value of i with respect to α	MAP (%)	Gain (%)
Anchor text	1	52.63	
Text extracted from source webpage	1 and 2	59.44	6.81
Text Extracted From Target webpage	3, 4 and 5	68.93	7.49
Collection of texts extracted in extraction of Hierarchical text	1, 2, 3, 4 and 5	74.67	5.74

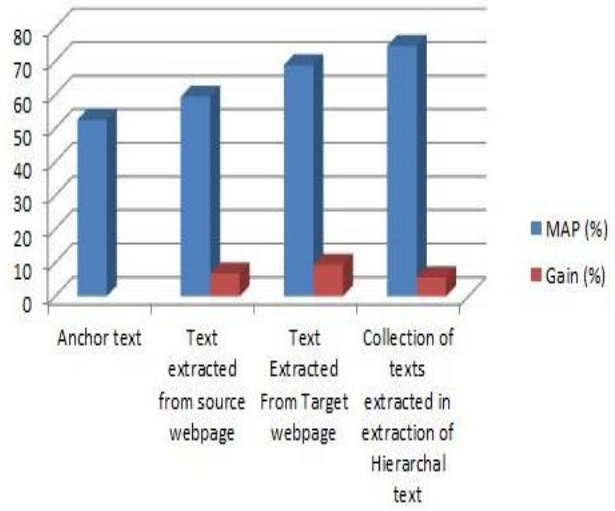

Figure 4. Mean Average Precision (MAP) and their gain values with respect to different texts.

For analyse the proposed work, the obtained result is compared with other ranking algorithms by introducing a system whose GUI snapshots are present in the Figures 5, 6, and 7. Figure 5 shows the front GUI page which take the input seed URL and query. After it crawl the seed URL and extract their respective outlink URLs and build their respective HTML trees which is shown in Figure 6. Thereafter this process of crawling is carried out on extracted URLs and finally rank value is calculated by applying proposed algorithm and PageRank algorithm [1] and Topical Distillation based on link evidence and webpage characteristics approach [19] which is shown in Figure 7.

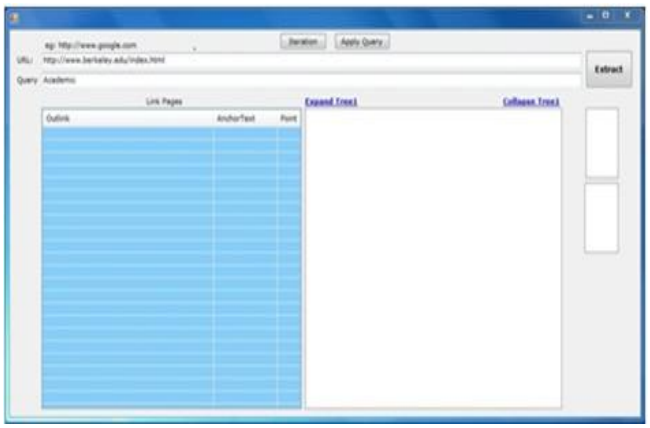

Figure 5. Snapshot 1: for given input, query and seed url of berkeley web portal i.e. http://www.berkeley.edu/index.hrml.

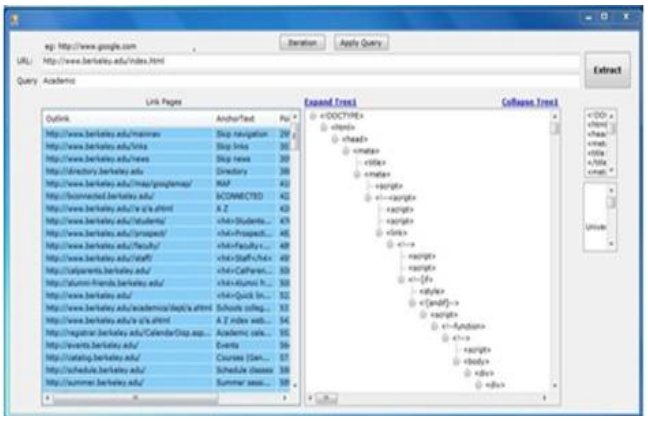

Figure 6. Snapshot 2: Crawl the seed webpage and extraction of its outlinks and build respected HTML tree.

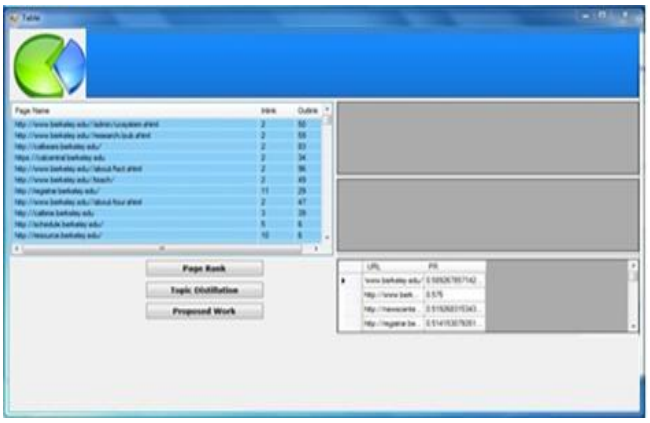

Figure 7. Snapshot 3: Count the number of inlinks, outlinks, and computation of respected similarity measures of rank values.

6. Discussion

Previous researches in the field of ranking were focused on the popularity of the target webpage which is calculated using different parameters such as inlinks,outlinks and works offline and independent to query. And in other researches which were query dependent and works online, popularity of target webpage is calculated using different texts extracted from source and target webpage.

In the proposed Hierarchical Rank algorithm, collection of different text sets as Hierarchical text is used with HTML tag importance parameter. This

importance parameter provides extra weightage to target webpage if keyword of query is found in corresponding HTML tag which has more importance.

There is a problem of ranking algorithms that is evenly distribution of rank to target webpage is solved by proposed algorithm. In the proposed algorithm rank is unevenly distributed to target webpage based on similarity of texts extracted from source and target webpage and query with respect to HTML tag importance parameter.

7. Conclusions

In this paper Hierarchical Rank algorithm is proposed for webpage ranking. In the proposed Hierarchical Rank algorithm, collection of different text sets as Hierarchical text is used with HTML tag importance parameter. This importance parameter provides extra weightage to target webpage if keyword of query is found in corresponding HTML tag which has more importance.In analysis of results, the experimental results of proposed algorithm are compared with other results of other two algorithms which show the improvement in terms of relevancy with respect to user's query. And gain in terms of MAP is also used in result analysis.

As the webpage rank value calculated in this research work is static in nature. In future proposed work, the content of webpage can be dynamic in nature. This can be fetched from databases based on user's query.Moreover,in future work accuracy and relevance can be improved by using dynamic content which is updated frequently.

References

[1] Brin S. and Page L., "The Anatomy of a Large Scale Hypertextual Web Search Engine," *Computer Network and ISDN Systems*, vol. 30, no. 1-7, pp. 107-11, 1998.

[2] Caverlee J., Webb S., Liu L., and Rouse W., "A Parameterized Approach to Spam-Resilient Link Analysis of the Web," *IEEE Transactions on Parallel and Distributed Systems*, vol. 20, no. 10, pp. 1422-1438, 2009.

[3] Chakrabarti S., Dom B., Raghavan P., and Rajagopalan S., "Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text," *in Proceedings of the 7th International Conference on World Wide Web 7*, Brisbane, pp. 65-74, 1998.

[4] Dubey H. and Roy B., "An Improved Page Rank Algorithm Based on Optimized Normalization Technique," *International Journal of Computer Science and Information Techniques*, vol. 2, no. 5, pp. 2183-2188, 2011.

[5] Duhan N., Sharma A., and Bhatia K., "Page Ranking Algorithms: A Survey," *in Proceedings IEEE International Conference on Advance Computing*, Patiala, pp. 1530-1537, 2009.

[6] Haveliwala T., "Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 784-796, 2003.

[7] Henzinger M., "Hyperlink Analysis for the Web," *IEEE Internet Computing*, vol. 5, no. 1, pp. 45-50, 2001.

[8] Hijikata Y., Hung B., Otsubo M., and Nishida S., "HITS Algorithm Improvement using Anchor-Related Text Extracted by DOM Structure Analysis," *in Proceedings of the ACM Symposium on Applied Computing*, Honolulu, pp. 1691-1698, 2009.

[9] Kleinberg J., "Authoritative Sources in a Hyperlinked Environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604-632, 1999.

[10] Kumar G., Duhan N., and Sharma A., "Page Ranking Based on Number of Visits of Webpages," *in Proceedings International Conference on Computer and Communication Technology*, Allahabad, pp. 11-14, 2011.

[11] Liu X., "An Improved HITS Algorithm Based on Page Query Similarity and Page Popularity," *Journal of Computers*, vol. 7, no. 1, pp. 130-134, 2012.

[12] Mohmmad A., Bidoki Z., and Yazdani N., "DistanceRank: An Intelligent Ranking Algorithm for Webpages," *Journal on Information Processing and Management*, vol. 44, pp. 877-892, 2007.

[13] Nie L., Davison B., and Qi X., "Topical Link Analysis for Web Search," *in Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, pp. 91-98, 2006.

[14] Noor S. and Bashir S., "Evaluating Bias in Retrieval Systems for Recall Oriented Documents Retrieval," *The International Arab Journal of Information Technology*, vol. 12, no. 1, pp. 53-59, 2015.

[15] Sharma D. and Sharma A., "A Comparative Analysis of the Page Ranking Algorithms," *International Journal of Computer Science and Engineering*, vol. 2, no. 8, pp. 2670-2776, 2010.

[16] Tao W. and Zuo W., "Query-Sensitive Self-Adaptable Webpage Ranking Algorithm," *in Proceedings International Conference on Machine Learning and Cybernetics*, Xi'an, pp. 413-418, 2003.

[17] Tyagi N. and Sharma S., "Weighted PageRanking Based on Number of Visits of links of Webpage," *International Journal of Soft*

Computing and Engineering, vol. 2, no. 3, pp. 441-446, 2012.

[18] Varadarajan R., Hristidis V., and Li T., "Beyond Single-Page Web Search," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 3, pp. 411-424, 2008.

[19] Wu M., Scholer F., and Turpin A., "Topic Distillation with Query Dependent Link Connections and Page Characteristics," *ACM Transactions on the Web*, vol. 5, no. 2, pp. 6:1-6:25, 2011.

[20] Xing W. and Ghorbani A., "Weighted Page Rank Algorithm," *in Proceedings Second Annual Conference on Communication Networks and Services Research*, Fredericton, pp. 305-314, 2004.

[21] Zhang Y., Xiao L., and Fan B., "The Research about Web Page Ranking Based on the A-PageRank and the Extended VSM," *in Proceedings of 5th IEEE International Conference on Fuzzy Systems and Knowledge Discovery*, Shandong, pp. 223-227, 2008.

Dilip Sharma is B.E.(CSE), M.Tech. (CSE) and Ph.D in Computer Engineering. He is Senior Member of IEEE, IEEE-CS, IEEE-WIEand Member of ACM, CSTA, USA and also life member of CSI, IETE, ISTE, IE, ISCA, SSI. He has published 72 research papers in International Journals /Conferences of repute and participated in 3 International/National conferences. He is consistently Conferred Significant Contribution Award by Computer Society of India in 47th and 48thCSI National Convention at Science City, Kolkata and Visakhapatnam, India. Presently he is working as Programme Coordinator (CSE) and Associate Professor in Department of Computer Engineering & Applications, GLA University, Mathura, U.P, India. He is Joint Secretary IEEE Uttar Pradesh Section and also Vice Chairman of Computer Society of India Mathura Chapter. His research interests are Web Information Retrieval and Software Engineering.

Deepak Ganeshiya received his B.Tech degree in Computer Science and Engineering from UPTU Lucknow, India in the year 2009 and M.Tech degree in Computer Science and Engineering from GLA University Mathura, India in the year 2014. During M.Tech his active area of research is Web information retrieval. He has three years of experience in the field of development of various e-governance projects.