IDENTIFYING SIMILAR WEB PAGES BASED ON AUTOMATED AND USER PREFERENCE VALUE USING SCORING METHODS

Gandhimathi K[1] and Vijaya MS[2]

[1]MPhil Research Scholar, Department of Computer science, PSGR Krishnammal College for Women, Coimbatore, India.
[2] Associate Professor, GR Govindarajulu School of Applied Computer Technology, Coimbatore, India.

ABSTRACT

A World Wide Web (WWW) is a system of interlinked hypertext documents accessed via internet. The web structure mining is based on the graph structure of hyperlinks and it extracts the useful information from structure of web data. Web structure mining aims to generate structural summary about web sites and web pages. Identifying the web community is one of the goal in web structure mining. This paper presents an alternate web community identification approach to detect the similar web pages in web community. The web community identification model is generated by extracting the attributes from the HTML source of input page. In the proposed work the preference value assigned by the user and automatically computed preference value for the given input page are used for analysis. The candidates are identified from the input page. Candidates are scored by three scoring methods such as normalized method, backlink analysis method and hyperlink analysis method and compared with preference value to find the similar pages in web community. The candidate pages with scores equal to preference value of the input pages are recognized as similar pages and hence it is found that the method produces scores equal to preference value which is an enhanced approach for identifying the similar pages. It is verified by using similarity analyzer tool by comparing the candidate pages with input page.

KEYWORDS

Automatic preference value, Similar pages, Scoring methods, User preference value, Web structure mining, Web community mining.

1. INTRODUCTION

WWW is becoming one of the most valuable resources for information retrieval and knowledge discovery. Retrieving information is the task of retrieving the relevant information from the web documents based on user query. A website is a collection of interlinked web pages that include a homepage residing at the same network location. Web mining is employed to automatically discover and extract information from the web documents and their services.

Web structure mining is the process of discovering structure information from the web and uses graph theory to analyze the connections. There are several goals for web structure mining such as ranking important web pages, finding the web communities, analysis of the web graph from macroscopic point of view, modelling and simulating the process of web graph generation. The main use of web structure mining is to extract previous unknown relationship between the web

DOI : 10.5121/ijdkp.2013.3603

pages. Web pages are connected to different locations using the structure component of hyperlink. Web structure mining categorizes the web pages and generates the information, like finding similar web pages and the relationship between different web sites [1]. The hyperlink analysis is used to find the path related information within the sites of the competitor links, connection throughout search engines and third party co-links. Web structure mining is based on the structure and composition of links in a given network of web sites and web structure mining provides insight into its structural characteristics.

Web community is a collection of web pages such that each member pages has more hyperlinks with in the communities than outside of the communities. Web community is based on the structure of hyperlinks which identify reasonable communities from a web graph constructed through query topic. Hyperlink information between two web pages conveys the intrinsic closeness and the relationship of content association.

Identifying the web communities is important to web user for information retrieval. Identification of communities on the web is significant for realistic applications include automatic web portals and focused search engines, content filtering and complement text-based searches. HITS algorithm [2] PageRank algorithm [3] and TrustRank algorithm [4] are common methods for web structure mining.

Various algorithms and methodologies have been developed for finding the web communities. In [5], the author has proposed a method for discovering web communities. In this method the related web pages are detected by the co-occurrence of hyperlinks on the pages which are acquired from a search engine. Hyperlinks contained in the acquired pages are extracted, and these pages are regarded as a new member of web community. A complete bipartite graph has been built with few URLs of initial community members and successfully discovered several genres of web communities without analysing the content of web page.

In [6], the author described the problem of extracting related community information from a large collection of web pages by performing hyperlink analysis. The dense bipartite graph (DBG) abstraction proposed for two purposes. First, it has extracted all the communities by mathematically abstracting the community as a DBG over a set of web pages from given web page collection. Next, the approach extracts related communities among the communities by abstracting related community set as a DBG over set of communities. The result demonstrated the meaningful community as well as related community structures.

The page similarity analysis and definition are based on hyperlink information among the web pages. The first algorithm, Extended Co-citation algorithm [7], is a co-citation algorithm that extends the traditional co-citation concepts. It is intuitive and concise. The second one, named Latent Linkage Information (LLI) algorithm [7], finds relevant pages more effectively and precisely by using linear algebra theories, especially the singular value decomposition (SVD) [8] of matrix, to reveal deeper relationships among the pages. The third one, named Equivalent Hyperlink algorithm [9] finds relevant pages more effectively and precisely by finding correlation between the two hyperlinks in terms of the anchor text and the content of the web page referred to for each hyperlink.

A search engine can be regarded as a resource for web data acquisition. In [5] [10], the authors proposed a method for discovering web communities from the data acquired from a search engine. In [11], the author described the search of bipartite graphs based on data acquired from a search engine. Search engines are used to find web pages about some keywords. Search engines followed the hyperlinks backward by attaching option “link:” to input URL, web pages that

contain hyperlinks to input URLs can be searched, which are called backlinks [12]. Hyperlinks to related web pages often co-occur, and backlink search used to find the related web pages.

In [13], the author has introduced user preference value as weights for the pages known to the user for finding of web communities by using the bipartite graph structure of hyperlinks. The methods proposed in [14] uses the algorithm described in [10] [13] and dealt with the finding algorithm of web communities in web structure mining. The method receives some URLs of user's known web pages, and proposes the user to find the candidates pages in the web community by using the structure of bipartite graph. The authors compared the values of the weights of output pages with input pages to detect the higher order similarity between the pages based on the structure of hyperlinks.

Figure 2.1 shows the sample network of pages and the links. Here centers are the input URL given by the user. Fans are the web pages that have hyperlink to centers. Candidates are the web pages that have backlinks from fans pages. The weights of user given pages i.e. input pages propagate through the hyperlinks and facilitated in finding the candidates of the pages in the web community.

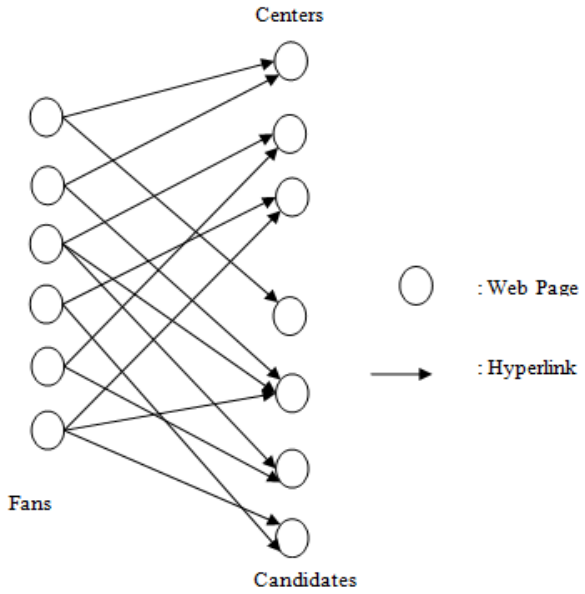

Figure 2.1 Sample network

This paper presents modified shimuzu algorithm wherein the preference value of input pages is generated automatically using term frequency and implemented using the real web pages collected from various domain such as educational institutions, financial institutions, organizations, legal institutions. The attributes such as topic, center URL, preference value, fan URL, Center and Fan (CF) value, candidate URL, Fan and Candidate (FC) value, hyperlinks count and backlinks count are extracted from the HTML source of web document. The method [15] first finds a community near the input URL and then finds broader communities. The proposed work compares the result of three scoring methods such as normalized method, backlink analysis method and hyperlink analysis method to obtain a best method for web community identification. The result of three scoring methods based on automated preference value is compared with the result of user preference value. The similarity analyzer tool is used to verify the text and HTML similarity of the three scoring methods.

The procedure for web community identification first acquires fans which have hyperlink to the input page by backlink search on a internet, from the HTML files of the acquired fans, all the hyperlinks are extracted. In order to find a new member of a web community, every hyperlinks contained in the acquired pages are extracted. The page which is pointed by the most frequent hyperlink is regarded as new member of the community. Web community is searched without analyzing content of the web pages.

2. PROPOSED WEB COMMUNITY IDENTIFICATION METHOD

The proposed work aims to find the web communities by using scoring methods along with the score of the candidates and to compare the candidate values with preference value of input page in order to generate top 10 candidate pages. The communities are identified for different web pages collected from various domains such as educational institutions, financial institutions, organizations and legal institutions.

The attributes are extracted from the HTML source code of a given web page. The input URL given by the user is a center page, the preference value is automatically extracted from the source code of the input page using term frequency. The preference value is the weight of input page and the output pages of the web community is computed based on weight of input page. The fan web pages that contain hyperlinks to input page can be searched from a internet, which are called backlinks.

The essential tasks carried out in this research work are computing the values of attributes, score value computation, comparing score value with preference and finding similar web pages. The architecture of the proposed system is shown in Figure 2.2.

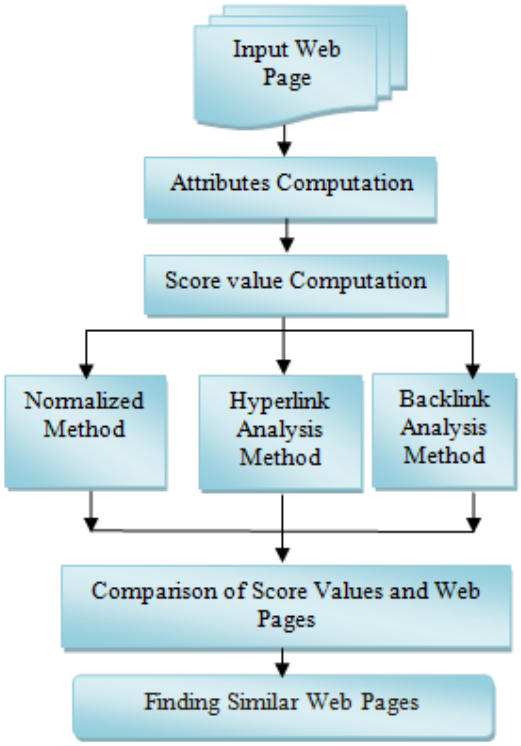

Figure 2.2 Proposed web community identification method

Related web pages are detected by the co-occurrence of hyperlinks contained in the fan pages and these pages are called as the candidate pages. The candidate page is an output of the web community. The attributes are computed based on source code of the input URL. Three types of scoring methods such as normalized method, backlink analysis method and hyperlink analysis method are employed to identify the candidate pages in order to detect the higher order similarity pages. The score values in the candidate pages in each method is compared with the preference value to find the most relevant page. Similarity analyzer tool is used for comparing the HTML code similarity and text similarity of the candidate pages with input page.

Center URL is an input URL given by user to find the similar pages in web community. The attributes such as topic, preference value, fan URL, CF value, candidate URL, FC value, backlinks count and hyperlinks count are extracted from HTML structure of input URL as described below

2.1. Topic

Title of the given URL is extracted from HTML structure; it states the title of the document. The title tag plays an essential role in search engine optimization. The syntax of title tag is <Title>Title of a web page </Title>. For example this tag <Title> WHO | World Health Organization </Title> returns “WHO | World Health Organization” as the topic of given input URL.

2.2. Preference Value

The preference values are obtained from the HTML structure of web pages using term frequency. Term frequency [16] [17] is used for computing the weight of a term that occurs in a web page. For example to find the term frequency for “The Educational Institutions” first it is converted into individual terms as “Educational”, “institutions” by removing the stop words such as the, of, and, for, etc.,. Then the frequency of each term is calculated and summed up to get the term frequency. If the term frequency is less than 10 then the preference value is set as 3, if the term frequency is greater than 9 and less than 15 then the preference value is set as 5, otherwise preference values is 7. The preference value is called as the weight of the input page. Term frequency value is calculated using the following formula,

$$tf_{ij} = \sqrt{\frac{n_{ij}}{\sum_k n_{kj}}}$$

2.3. Fans URL

The pages having the hyperlink to input page are fan pages. The hyperlink tag starts with "<a", and includes a reference "href="URL″>", is the start of the link. The text between the hyperlink tag and its corresponding "closing" tag (“") is called anchor text and the respective URL is called fan URL. Thus for each input page several fan URLs can be extracted. For example “http://www.wpro.who.int/vietnam/topics/cholera/en/” is the fan page backlink of the input page “http://www.who.int/tobacco/global_report/2013/en/ index.html”.

2.4. CF value

The fan pages are those pages which have links from input page. If the input page is referred back by a fan page then the CF value returns 1 otherwise 0. For example $X = \{x_1, \ldots, x_l\} \in P$ be the

centers, Y ={$y_1, \ldots, y_m$ } ∈P be the fans, and Z = {$z_1, \ldots, z_n$ } ∈P be the candidates of pages in the web community where P is the set of whole web page that define as A (p1, p2) if p2 has a hyperlink to p1 the value will be 1 otherwise 0.

2.5. Candidate URL

The pages that have hyperlink from the fan page are candidate pages. The group of candidate pages forms a community and it is an output of the web community mining. Candidate URLs can be extracted from HTML source code of a fan pages. Thus for each fan page several candidate URLs can be extracted. For example, "http://www.who.int/topic/ research/en/index.html" is the candidate page extracted from the hyperlink of fan page "http://www.wpro. who.int/vietnam/topics/cholera/en/".

2.6. FC value

The candidate pages are those pages which have links from fan pages. If the fan page is referred back by candidate page then the FC value returns 1 otherwise 0. The fans and the candidates are defined as $Y = \{y_j | \exists\, x_i \in X, A(y_j, x_i) = 1\}$, $Z = \{z_k | \exists y_j \in Y, A(y_j, z_k) = 1, z_k \notin X\}$.

2.7. Backlinks Count

The web page is defined as pi, Nin (pi) is the number of backlinks of the web page pi. It counts number of hyperlink from fan page to input page.

2.8. Hyperlinks Count

It counts number of hyperlink from fan page to candidate page. Nout (pi) is the number of hyperlinks of the web page pi.

Thus a list of 8 attributes is extracted from the HTML source code based on input URL is shown in Figure 2.3.

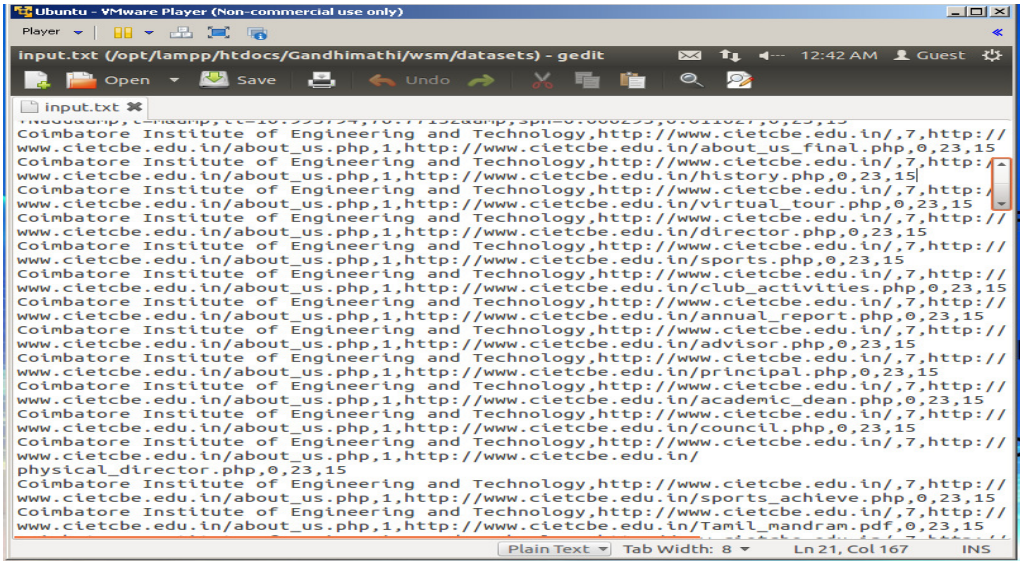

Figure 2.3 Extracted Attributes

3. SCORING METHODS

The scoring methods are used to score the candidate pages in web community to detect the similar pages. There are three scoring methods such as normalized method, backlink analysis with FC value and hyperlink analysis with FC value are used to find the relevant pages from candidate pages of web community. The scoring methods are based on mutual linkage information which is used to find the relevant pages in the web community related to input page given by the user using preference value. $X \rightarrow \{1, \ldots, N_{pref}\} \in N$ be the preference value N_{pref} is the maximum value of preferences.

The algorithm proposed to find the score of fans and candidates using center URL and its preference value. The fan value is calculated using the following formula.

$$score_Y(y_j) = \sum_i pref(x_i)A(x_i, y_j) \qquad \ldots\ldots\ldots\ldots\ldots(1)$$

The $score_Y(y_j)$ is score of the fan page. The $pref(x_i)$ is preference value of the center, (x_i, y_j) returns 1 if y_j have hyperlink to x_i otherwise return as 0.

$$score_Z(z_k) = \sum_j score_Y(y_j)A(z_k, y_j) \qquad \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

The $score_Z(z_k)$ is a score of candidate page. The $score_Y(y_j)$ is a score of fan page, (z_k, y_j) retum 1 if y_j have hyperlink to z_k otherwise return as 0. The score of the page x_i in the centers is

$$score_Z(x_i) = \sum_j score_Y(y_j)A(x_i, y_j)$$

$$= \sum_j \left(\sum_k pref(x_k)A(x_k, y_j) \right) A(x_i, y_j)$$

$$\geq Nin(x_i)pref(x_i) \qquad \ldots\ldots\ldots\ldots\ldots(3)$$

$$score_Z(x_i) \neq pref(x_i)$$

The $score_Z(x_i)$ is score of the candidate based on input page. $Pref(x_i)$ is the preference value of the input page. The weights of the centers propagate the same value for each hyperlinks or backlinks, and the weight of each page in the fans or candidates is calculated by sum of those values. So the output of the score is cannot compare the score of candidates with preference of centers, because there is a difference between the score values and preference value. The values of weights of candidates are efficient to sort the candidates, but not efficient for user to compare the center and the candidates. Moreover, the score of candidates is used for sorting the candidates to find the most relevant pages. The following three scoring methods are used to find the candidate score value and user to compare with preference of centers based on above formula.

Normalized Method

The formula to normalize the score value of the candidates with preference value is,

$$score_Z^{(1)}(z_k) = \frac{N_{pref}}{\max_i score_Z(z_i)} score_Z(z_k) \qquad \ldots\ldots\ldots\ldots\ldots(4)$$

Backlink Analysis Method

The score value of fan page is calculated using preference of the center, number of hyperlinks from the center it propagate to the fan and CF value.

$$score_Y^{(2)}(y_j) = \sum_i \frac{pref(x_i)}{N_{in}(x_i)} A(x_i, y_j) \qquad \text{.....(5)}$$

The candidate score value is calculated by score value of the fan and CF value using the formula,

$$score_Z^{(2)}(z_k) \sum_j score_Y^{(2)}(y_j) A(z_k, y_j) \qquad \text{.....(6)}$$

Hyperlink Analysis Method

To find the score of the candidate divide the score of the fans by the number of hyperlink from the fans it propagates to the candidates, and calculated with FC value.

$$score_Z^{(3)}(z_k) = \sum_i \frac{score_Y(y_j)}{N_{out}(y_j)} A(z_k, y_j) \qquad \text{.....(7)}$$

For the normalized scoring method, the candidates are the same order to equation (3). In the backlink analysis method or hyperlink analysis method, if there are no pages that have hyperlinks to more than one pages in the center, the score of candidate is equal to the preference of the input page. The preference value that is intuitively given by the user does not always have consistency. For example, if two pages in the center have common fans, these pages will have same score, but the user may give the different values of preference. In spite of the above problem, it is important that the scores of candidates have similar values to the preference of the center. The above three methods are used to find the candidate scores which results in finding similar pages to detect web community. This finds the candidates of web community and the pages similar to input page by comparing the score of candidates with preference value. For example, for a given input page the proposed model generates and displays pages, that the user may be interested.

3.1. Web Page Similarity Analyzer Tool

The scoring methods are based on structure of link analysis. The score of the candidates are compared with preference value. The candidate pages are compared with the input page using similarity analyzer. The similarity analyzer measures the HTML similarity and the text similarity. HTML similarity value can be used to identify pages belonging to the same site or network and text similarity value is used to find the duplicated or near identical text content. The result produced by similarity analyzer tool is shown in Figure 3.1.

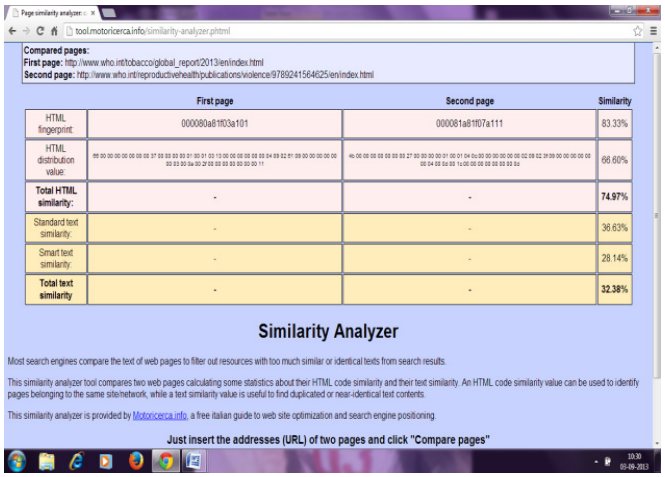

Figure 3.1 similarity analyzer

4. EXPERIMENTS AND RESULTS

Two independent experiments have been carried out for finding similar pages. First experiment is based on user preference value and the second is based on automatic preference value. These experiments have been carried out by implementing scoring methods for web community identification. The methods namely, normalized method backlink analysis method and hyperlink analysis method are employed here.

Several web pages are collected from various domains such as educational institutions, financial institutions, organizations, legal institutions and a data set is created. For each domain 20 web pages are collected and totally 80 web pages are used for experimenting web community identification. For each web page 8 attributes such as topic, preference value, fan URL, CF value, candidate URL, FC value, backlinks count and hyperlinks count are extracted for computing the score values. The preferences values are set as 7, 5, and 3 for input URL. Preference value is weight of the input page and it propagates through the hyperlink and backlink. It is used to find the score of the candidate to detect the pages that are similar to the input page.

The attributes are extracted from the HTML source code. These attributes are used for finding the score of candidates using three scoring methods. Normalized method normalizes the score of the candidate using preference value. In backlink analysis method, candidate score is calculated based on its backlinks count and FC value. Using hyperlink method, the score of candidate pages are computed based on hyperlinks count and FC value.

In the first experiment, the preference values 7, 5 and 3 are assigned by user for the given input page. The scoring methods such as normalized method, backlink analysis method, hyperlink analysis method are used to find the score of the candidate.

The preference value given by the user is not always consistent which propagates to change according to the user. When a page has high preference with more backlinks, the user gives low preference to those pages. In such case, both normalized method and hyperlink analysis method produce high scores for the candidate pages. In backlink analysis method the score of the fan is divided by the number of fans, so that it results in low score value. The result of the scoring methods such as normalized method, backlink analysis method, hyperlink analysis methods are shown in TABLE I.

In the second experiment, the preference value is calculated automatically using term frequency for finding web community. The term frequency is calculated from the HTML source code of input page based on title keyword. The value 7, 5 and 3 will be generated automatically as preference value.

Backlinks of the input page are fan pages, which are generated by internet in order to follow hyperlink backward. HTML files of the searched fans are acquired through the internet, along with all the hyperlinks contained in the files are extracted. All the hyperlinks are sorted in the order of frequency. Therefore, the URL of the page is added as a new member of center as candidates. The above two steps such as generating fans and generating candidates are repeatedly applied until there are few fans which refer all the members of center. Resulted candidates are regarded as a web community.

The top 10 ranked candidate score values from three scoring methods are compared with preference value of the center. User can compare the candidate score value to find the similar

range of preference value. The scoring methods are used to sort the candidates and to find the similar pages in web community. Comparison of score values with automated preference value is shown in TABLE I.

TABLE I – Comparison of score values with automated preference value and user preference value

Rank	Preference		Normalized method		Backlink Analysis Method		Hyperlink Analysis Method	
	Automatic	User	Automatic	User	Automatic	User	Automatic	User
1	7	3	7.00	5.80	7.00	2.20	13.90	9.00
2	7	3	6.30	5.30	7.00	2.10	12.90	7.40
3	7	3	5.00	5.00	7.00	1.40	12.40	6.80
4	7	3	5.00	3.96	7.00	0.90	12.10	6.20
5	7	3	4.50	3.72	7.00	0.30	11.00	6.10
6	7	3	4.20	3.00	7.00	0.10	5.80	5.80
7	7	3	3.00	3.00	7.00	0.08	5.60	5.60
8	7	3	2.90	1.06	7.00	0.03	2.20	2.20
9	7	3	2.60	1.03	6.60	0.02	1.10	1.10
10	7	3	2.50	1.01	6.50	0.01	0.90	0.90

From the above results it is concluded that the calculated score value equal to its preference value is desirable. The score results shows there is a vast difference in score values and unable to compare the preference value with score values while finding the similar pages. In automated preference method the score of candidate pages shown by the normalized method is smaller than the preference values. The candidate score shown by the backlink analysis method is almost similar to preference values. The score of candidate pages shown by the hyperlink analysis method is higher than the preference values of input page. The automated preference value method is better than the user preference value when comparing the preference value with three scoring methods. The resulted top 10 pages based on three scoring methods are compared with the input page by using similarity analyzer tool.

Similarity analyzer tool is used for analyzing the HTML and the text similarity of the top 10 candidates. It compares the candidate pages in each method with input pages to find the similarity value. HTML similarity and text similarity of the normalized method, backlink analysis method and hyperlink analysis method are shown in TABLE II.

The average and comparative result of the HTML similarity and text similarity in three scoring methods is shown in TABLE III.

TABLE II - HTML and text similarity of all the three Scoring Methods

Normalized Method		Backlink Analysis Method		Hyperlink Analysis Method	
HTML	Text	HTML	Text	HTML	Text
74.97%	32.38%	91.11%	45.21%	48.65%	16.56%
71.19%	27.29%	90.50%	47.35%	42.15%	27.68%
69.55%	33.08%	88.81%	45.02%	51.31%	17.23%

68.65%	31.43%	87.58%	43.37%	47.66%	13.39%
66.93%	35.84%	86.96%	45.08%	46.93%	17.03%
66.02%	27.37%	86.27%	42.36%	42.32%	8.53%
62.14%	23.51%	84.17%	45.03%	45.41%	7.01%
61.40%	27.19%	83.39%	35.39%	40.93%	19.10%
56.94%	22.82%	80.45%	30.43%	35.46%	15.99%
59.37%	20.29%	78.39%	38.18%	36.39%	0.98%

TABLE III – Average Result of three Scoring Methods

Normalized Method (%)		**Backlink Analysis Method (%)**		**Hyperlink Analysis Method (%)**	
HTML	**Text**	**HTML**	**Text**	**HTML**	**Text**
65.71	31.78	85.76	41.74	43.82	14.35

The overall comparison of HTML similarity and text similarity in the normalized method, backlink analysis method and hyperlink analysis method is shown in Figure.4.1.

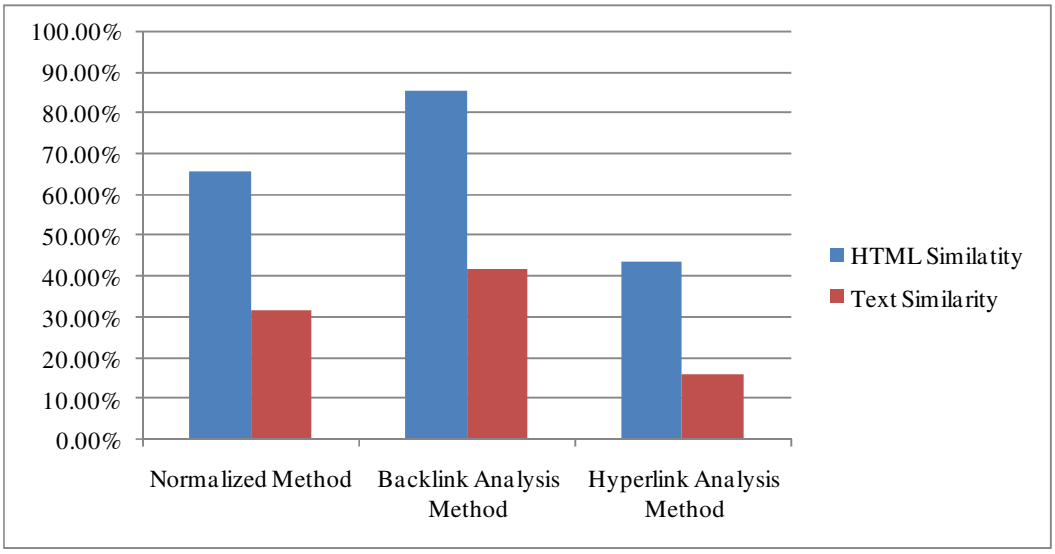

Figure.4.1 Comparison of web page similarity

The overall percentage shown by first method for the HTML similarity is 65.71% and text similarity is 31.78%. The overall percentage shown by second method for the HTML similarity is 85.76% and text similarity is 41.74%. The overall percentage shown by third method for the HTML similarity is 43.82% and text similarity is 14.35%. Hence, the candidate score are similar preference value and candidate pages are highly similar to input page in backlink analysis method when compared to other two methods.

5. CONCLUSIONS

This paper described modified approach for identifying the web communities using preference value of a given web page. Three scoring methods such as normalized method, backlink analysis method, hyperlink analysis method have been used to compute the score of candidate pages. In the backlink analysis method, the value of candidate score has equal value to its preference value. The score of candidate pages in backlink analysis method depends on the relationship between number of backlinks and preference of the center.

The similarity analyzer tool has been used to compare the input page with candidate page. The backlink analysis method yield better results than the normalized method and hyperlink analysis methods, in generating similar pages and it is verified by using similarity analyzer tool. Also in this work the preference values of input pages are computed automatically using term frequency since the preference that is intuitively given by the user may not be consistent.

As a scope for future work the web community detection model can be enhanced by giving the preference value for the candidates. Also the model can be further enhanced by using the content of fan pages and candidate pages in order to detect the similar pages for web community mining.

REFERENCE

[1] Zhiguo Gong, "Web Structure Mining: An Introduction". Proceedings of the 2005 IEEE International Conference on Information Acquisition June 27 - July 3, 2005, Hong Kong and Macau, China.

[2] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "The Web as a Graph; Measurements, Models, and Methods," Proc. of the 5th Annual Int. Conf. on Computing and Combinatorics, Lecture Notes in Computer Science, Vol.1627, 1999.

[3] L. Page, S. Brin, R. Motwani, and T.Winograd, "The Page Citation Ranking: Bringing Order to the Web," Technical Report, Stanford University, 1998.

[4] Z.Gyongyi, H. Garcia-Molina, and J. Pedersen, "Combating Web Spam with TrustRank," Technical Report, Stanford University, 2004.

[5] T. Murata, "Discovery of Web Communities Based on the Co-occurrence of References", Proc. of DS2000, pp.65-75, 2000.

[6] Jingyu Hou and Yanchun Zhang, "Effectively Finding Relevant Web Pages from Linkage Information", August 2003.

[7] B.N. Datta, "Numerical Linear Algebra and Application". Brooks / Cole Publishing, 1995.

[8] Simon Courtenage and Steven Williams, "Finding Relevant Web Pages Through Equivalent Hyperlinks", University of Westminster, Apr.2004.

[9] P.Krishna Reddy and Masaru Kitsuregawa, "An Approach to Find Related Communities Based on Bipartite Graphs" Institute of Industrial Science, The University of Tokyo, 2001.

[10] T. Murata, "Finding Related Web Pages Based on Connectivity Information from a Search Engine," Poster Proc. of the Tenth Int.World Wide Web Conf. (WWW10), 2001.

[11] Tsuyoshi Murata, "Graph Mining Approaches for the Discovery of Web Communities" National Institute of Informatics 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 2003.

[12] Michael chau, Body shiu, Ivy chan, Hsinchun chen, "Automated Identification Web Communities for Business Intelligence Analysis", 2007.

[13] H. Shimizu, "A Study on Web Mining Based on Web Communities" Graduation Thesis, Hokkaido University, 2006.

[14] Takeshi Yoshikawa and Hidetoshi Nonaka, "Finding Communities Using User Preference in Web Structure Mining" Journal of Advanced Computational Intelligence and Intelligent Informatics, 2011.

[15] Jiayuan Huang1, 2, Tingshao Zhu2, and Dale Schuurmans, "Web Communities Identification from Random Walks", University of Waterloo,2006.

[16] Santhana Lakshmi V and Vijaya M.S, "The SVM Based Interactive Tool for Predicting Phishing Websites", in proceedings of international Journal of Computer Science and Information Security,Vol.9 No.10, Oct 2011.

[17] Kavitha S and Vijaya MS, "Classifying Web Pages using Support Vector Machine" Elixir Comp. Sci. & Engg. 55 (2013) 12825-12829.

AUTHORS

K. Gandhimathi is pursing Master of Philosophy in Computer Science in PSGR Krishnammal college for women under the guidance of MS.Vijaya. Her research interests are data mining, web mining and social networks.

MS. Vijaya is presently working as Associate Professor in GR Govindarajulu School Of Applied Computer Technology, PSGR Krishnammal college for women, Coimbatore, India. She has 22 years of teaching experience and 8 years of research experience. She has completed her doctoral programme in the area of Natural Language Processing. Her areas of interest include Data Mining, Support Vector Machine, Machine learning, Pattern Recognition, Natural Language Processing and Optimization Techniques. She has presented 22 papers in National conferences and she has to her credit 17 publications in International conference proceedings and Journals. She is a member of Computer Society of India, International Association of Engineers (Hong Kong), International Association of Computer Science and Information Technology (IACSIT – Singapore). She is also a reviewer ofInternational Journal of Computer Science and Information Security.