International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

An Enhanced Web Mining Technique for Image Search using Weighted PageRank based on Visit of Links and Fuzzy K-Means Algorithm

Rashmi Sharma[1], Kamaljit Kaur[2]

[1]Student, M. Tech in computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib - 140406, Punjab, India

[2]Assistant Professor, Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib - 140406, Punjab, India

Abstract: ***Web mining is growing very rapidly. It extracts the rich information in the form of audio, video, content, links and images. The World Wide Web contains huge number of web pages and information available within web pages. When a user gives a query to the search engine, it generally returns a large amount of information in response to user's query. To retrieve required information from Web pages, web mining performs various techniques. This paper proposed the hybrid technique of Weighted PageRank based on Visit of Links and Fuzzy K-Means algorithms which are applied on the search result. This paper described Fuzzy K-Means algorithm is used to group the given data-set into clusters and Weighted PageRank is used to re-rank the data according to the visit of links in the web pages. We have fetched the relevant information such as image links, images and total hyperlinks from a URL using the hybrid approach. Furthermore, we have done analysis on hybrid algorithms by applying valid parameters like execution time, recall, precision and f-measure.***

Keywords: PageRank, Weighted PageRank, PageRank based on Visit of Links, Weighted PageRank based on Visit of Links, K-Means and Fuzzy K-Means

1. Introduction

Web mining is an application of data mining. It defined as the process of extracting useful information from World Wide Web data. It employs text, audio, video, contents and images from web pages. Two different approaches are taken to define the web mining. (1) "Process-centric view" which defines web mining as sequence of task. (2) "Data-centric view" which defines web mining in terms of the types of web data that is being used in the mining process [7].

Web mining can be broadly divided into three distinct categories, according to the types of data to be mined, Web Content Mining, Web Structure Mining and Web Usage Mining[11]. Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page contents. Content mining is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query [8]. Web usage mining is the process of extracting useful information from server logs i.e. users' history. Web usage mining is the process of finding out what users are looking for on Internet. Web usage mining process involves the log time of pages. The world's largest portal like yahoo, msn etc., needs a lot of insights from the behavior of their users' web visits [8] and Web structure mining, one of three categories of web mining for data, is a tool used to identify the relationship between Web pages linked by information or direct link connection. This structure data is discoverable by the provision of web structure schema through database techniques for Web pages [8].

In this paper our approach will extend to image web links, the proposed work has done on web links search and including hyperlinks and image links using Weighted PageRank based on visit of links along with Fuzzy K-Means clustering. In this research the problem is formulated on efficient search of Web Links along with images, Weighted PageRank and Fuzzy K-means will be followed in link and image search and at the end various parameters will be studies. This approach holds simultaneously two problems of link and image search. The algorithm provide the relevant links which the user wants as sometimes when user requires the image some non relevant images are also got extracted which is poor knowledge discovery in data mining and inaccurate approach.

The rest of this paper is organized as follows: a brief summary of related work is given in section 2. Section 3 described the proposed algorithms in detail. The methodology of proposed work is illustrated in section 4. Section 5 demonstrated the results, screenshots of proposed method and experiments applied on hybrid technology and illustrate a general conclusion in section 6.

2. Related Work

Brin and Page [1, 6] developed PageRank algorithm at Stanford University based on the hyper link structure. PageRank algorithm is used by the famous search engine, Google. PageRank algorithm is the most frequently used algorithm for ranking billions of web pages. During the processing of a query, Google's search algorithm combines precomputed PageRank scores with text matching scores to obtain an overall ranking score for each web page. Functioning of the Page Rank algorithm depends upon link structure of the web pages. The PageRank algorithm is based on the concepts that if a page surrounds important links

towards it then the links of this page near the other page are also to be believed as imperative pages. The Page Rank imitate on the back link in deciding the rank score. Thus, a page gets hold of a high rank if the addition of the ranks of its back links is high. A simplified version of PageRank is given as:

$$PR(u) = (1-d) + d\sum_{v\varepsilon B(u)}\left(\frac{PR(v)}{Nv}\right)$$

Where u represents a web page, $B(u)$ is the set of pages that point to u, $PR(u)$and $PR(v)$ are rank scores of page u and v respectively, Nv indicates the number of outgoing links of page v, d is a damping factor.

Gyanendra Kumar et. al. [4, 6] proposed a new algorithm in which they considered user's browsing behaviour. As most of the ranking algorithms proposed are either link or content oriented in which consideration of user usage trends are not available. In this paper, a page ranking mechanism called Page Ranking based on Visits of Links is being devised for search engines, which works on the basic ranking algorithm of Google, i.e. PageRank and takes number of visits of inbound links of web pages into account. This concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behavior, which reduce the search space to a large scale. In this paper as the author describe that in the original PageRank algorithm, the rank score of page p, is evenly divided among its outgoing links or we can say for a page, an inbound links brings rank value from base page, p. So, he proposed an improved PageRank algorithm. In this algorithm we assign more rank value to the outgoing links which is most visited by users. In this manner a page rank value is calculated based on visits of inbound links. The modified version based on VOL is given as:

$$PR\,u = (1-d) + d\sum_{v\varepsilon B(u)}\left(\frac{L_u PR(v)}{TL(v)}\right)$$

Notations are:

- d is a dampening factor ,
- u represents a web page,
- $B(u)$ is the set of pages that point to u,
- $PR(u)$ and $PR(v)$ are rank scores of page u and v respectively,
- L_u is the number of visits of link which is pointing page u from v.
- $TL(v)$ denotes total number of visits of all links present on v.

Wenpu Xing et. al. [3, 6] discussed a new approach known as weighted pagerank algorithm (WPR). This algorithm is an extension of PageRank algorithm. WPR takes into account the importance of both the inlinks and the outlinks of the pages and distributes rank scores based on the popularity of the pages. WPR performs better than the conventional PageRank algorithm in terms of returning larger number of relevant pages to a given query.

According to author the more popular web pages are the more linkages that other web pages tend to have to them or are linked to by them. The proposed extended PageRank algorithm–a Weighted PageRank Algorithm–assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its outlink pages. Each outlink page gets a value proportional to its popularity (its number of inlinks and outlinks). The popularity from the number of inlinks and outlinks is recorded as $W_{in(v,u)}$ and $W_{out(v,u)}$, respectively. $W_{in(v,u)}$ is the weight of $link\,(v,u)$ calculated based on the number of inlinks of page u and the number of inlinks of all reference pages of page v. $W_{out(v,u)}$ is the weight of $link\ (v,u)$ calculated based on the number of outlinks of page u and the number of outlinks of all reference pages of page v. Considering the importance of pages, the original PageRank formula is modified as:

$$PR(U) = (1-d) + d\sum_{v\varepsilon B(u)} PR(V)W^{in}(U,V)W^{out}(U,V)$$

Syed Thousif Hussain et al. [5] have proposed the approach which is to generate a large number of images for specified object class. This approach is to employ text, metadata and visual features and to use to gather many high quality images from the web. Candidates images are obtained by text based web search. The web page and the images are downloaded. The task is to remove irrelevant images and to re-rank. First, the images query page is down-loaded. Second, it extracts images URL from downloaded page and place it in the database then ranking is done based on text surrounding and metadata features. SVM (Support Vector Machine) and Naive bayes classifier algorithm are compared for ranking. The top ranked images are used as training data and an SVM visual classifier is learned to improve re-ranking. The principal idea of the overall method is in combining text or metadata or visual features in order to achieve a completely automatic ranking of images.

Preeti Chopra et al. [7] have purposed web structure mining algorithms like pagerank algorithm, weighted pagerank algorithm, weighted content pagerank algorithm (WCPR), HITS etc. This paper analyzed their strengths and limitations and provides comparison among them. In order to achieve this goal, this paper uses the concept of web mining. Web mining is used to categorize users and pages by analyzing the users' behavior, the content of the pages, and the order of the URLs that tend to be accessed in order .Web structure mining is defined as the process of analyzing the structure of hyperlink using graph theory. This paper may be used as a reference by researchers when deciding which algorithm is suitable and also try to overcome from the problem that particular algorithms have.

Gurjit Singh et al. [10] have studied that clustering is an essential task in Data Mining process which is used for the purpose to make groups or clusters of the given data set based on the similarity between them. This paper is intended to give the introduction about K-means clustering and its algorithm. The experimental result of K-means clustering and its performance in case of execution time is discussed here. But there are certain limitations in K-means clustering algorithm such as it takes more time for execution. So in order to reduce the execution time this paper is using the Ranking Method. And also shown that how clustering is performed in less execution time as compared to the traditional method. This work makes an attempt at studying

the feasibility of K-means clustering algorithm in data mining using the Ranking Method. Modifications in hard K-means algorithm such that algorithm can be used for clustering data with categorical attributes. To use the algorithm for categorical data modifications in distance and prototype calculation are proposed. To use the algorithm on numerical attribute values, means is calculated to represent centre, and Euclidean distance is used to calculate distance.

Rashmi Rani et al. [12] studied that **i**nformation available on the WWW, users' get easily lost in rich hyper structure. In this paper, finding the content of the web and retrieving the users' interest and need from their behaviour has become important .Web mining is used to cauterized user and pages by analyzing the user's behaviour, content of pages, order of the URLs, Two page ranking algorithms, HITS and PageRank. In this paper the relevancy values for the query produced by Page Rank and WPR using different page set and finally search result list is reranked by updating the existing page rank values of a page. The proposed work result WPR performs better than Page Rank and reduced search time and important pages are tending to move upwards in the result list.

Supreet Kaur et al. [13] have studied that Clustering is an essential task in Data Mining process which is used for the purpose to make groups or clusters of the given data set based on the similarity between them. This paper is intended to give the introduction about K-means clustering and its algorithm. The experimental results of K-means clustering and its performance in case of execution time are discussed here. But there are certain limitations in K-means clustering algorithm such as it takes more time for execution. So in order to reduce the execution time we are using the weighted page rank with k means clustering and also shown that how clustering is performed in less execution time as compared to the traditional method. This work makes an attempt at studying the feasibility of K-means clustering algorithm in data mining using the weighted page rank with k means clustering. K means with page rank algorithm gave results with better result set of various numbers of data-sets. In this research the work is going on k means clustering of database with weighted page content rank algorithm.

3. Proposed Approach

Here we proposed hybrid approach. In this hybrid technique two well known web mining algorthims is used such as Weighted PageRank based on Visit of Links and Fuzzy K-Means. This proposed method is implemented in two steps: Firstly Fuzzy K-Means is used to group the data set into clusters and secondly Weighted PageRank is implemented on cluster to rank the data in it according to the visit of links. These two algorithms are explained in detail as following:

3.1 Fuzzy K-Means

The fuzzy k-means clustering algorithm partitions data points into k clusters Sl (l = 1,2, …, k) and clusters Sl are associated with representatives (cluster center) Cl. The relationship between a data point and cluster representative is fuzzy. The major process of FKM is mapping a given set of representative vectors into an improved one through partitioning data points. It begins with a set of initial cluster centers and repeats this mapping process until a stopping criterion is satisfied. It is supposed that no two clusters have the same cluster representative. In the case that two cluster centers coincide, a cluster center should be perturbed to avoid coincidence in the iterative process. The fuzzy k-means clustering algorithm is now presented as follows[9].

Algorithm: There are following steps of algorithm give as:

1. **Pre-processing Phase**

Step 1: Select informative genes by using Entropy filtering approach. The Genes with low entropy value is re-moved.

Step 2: Fuzzify feature vector values using S-shaped membership function or Z-shaped membership function.

2. **Clustering Phase**

Input: K-No of clusters.

n- Total number of Genes.

m- Number of samples.

W_{lower}, W_{upper} and threshold (ε).

Output**:** K-Gene clusters.

Step 1: Randomly assign each object into exactly one lower Approximation C_k, the objects also belongs to upper approximation C_k of the same cluster. Boundary region isC_k^B.

Step 2: Compute cluster centroids.

$i = 1, 2 \ldots, n,\ j = 1, 2 \ldots m \text{ and } h = 1, 2, \ldots, k$

If $C_K^B = \overline{C_k}.\underline{C_k} \neq \emptyset$

$$Z_k = \left(W_{lower} \times \frac{\sum X \varepsilon \underline{C_k} X_i}{|\underline{C_k}|}\right) + \left(W_{lower} \times \frac{\sum X \varepsilon \underline{C_k} X_i}{|\overline{C_k} - \underline{C_k}|}\right)$$

Else

Compute new centroids,

$$Z_k = \sum x \varepsilon c \ \frac{X}{|C_k|}$$

End

Step 3: Find Similarity Si, Here, i-Represents genes,

$$S_i(\hat{X}, \hat{Z}) = 1 - \frac{\sum_{j=1}^{n} |\hat{x}_{ij} - \hat{z}_{hj}|}{\sum_{j=1}^{n} (\hat{x}_{ij} + \hat{z}_{hj})}$$

Step 4: Compute $Pi = \frac{Max_{Si}}{Min_{Si}}$

and normalize the Pi values

If $\geq (\varepsilon)$, insert i^{th} object in K^{th}Cluster.

Step 5: Update centroids. Repeat the steps 2 to step 5, until New centroid = Old centroid.

3.2 Weighted PageRank based on Visit of Links

In Weighted PageRank algorithm, we assign more rank value to the outgoing links which is most visited by users and received higher popularity from number of inlinks. The user's browsing behavior can be calculated by number of hits (visits) of links. The modified version based on WPR (VOL) is given as [6].

$$WPR_{vol}(u) = (1 - d) + d \sum_{v \varepsilon B(u)} \frac{L_u WPR_{Vol}(v) W_{in}(v,u) TL(v)}{TL(v)}$$

Algorithm: The various steps of the proposed algorithm are given below [6].

Step 1: Finding a Website: Find a website which has rich hyperlinks because WPR (VOL) methods rely on the web structures.

Step 2: Building a Web Map: Then generate the web map from the selected website

Step 3: Calculate W^{in} (v,u): Then calculate the W^{in} (v,u) for each node present in web graph by applying the equation as below.

$$W^{in}(m,n) = \frac{I_n}{\sum_{p \varepsilon R(m)} I_p}$$

Where

- $W^{in}(v,u)$ is the weight of link (v,u) calculated based on the number of inlinks of page u and the number of inlinks of all reference pages of page v.
- In and Ip are the number of incoming links of page n and page p respectively.
- $R\ (m)$ denotes the reference page list of page m.

Step 4: Apply proposed formula: Now calculate the PageRank value of the nodes present in web graph by using the proposed formula

$$WPR_{vol}(u) = (1-d) + d \sum_{v \varepsilon B(u)} \frac{L_u WPR_{Vol}(v) W_{in}(v,u) TL(v)}{TL(v)}$$

Notations are:

- d is a dampening factor ,
- u represents a web page,
- $B(u)$ is the set of pages that point to u,
- $WPR_{vol}\ (u)$ and $WPR_{vol}\ (v)$ are rank scores of page u and v respectively,
- L_u is the number of visits of link which is pointing page u from v.
- $TL(v)$ denotes total number of visits of all links present on v.

Step 5: Repeat by going to step 4: final step will be used recursively until the values are to be stable.

4. Methodology

It describes the web based image search as user extract the relevant information in the form of image links, images and hyperlinks from the URL and compute the number of visits using Weighted PageRank from the URL. At the end we compute various parameters such as Execution Time, Recall, Precision and F-Measure. The whole process of proposed work as shown in the form of flow chart:

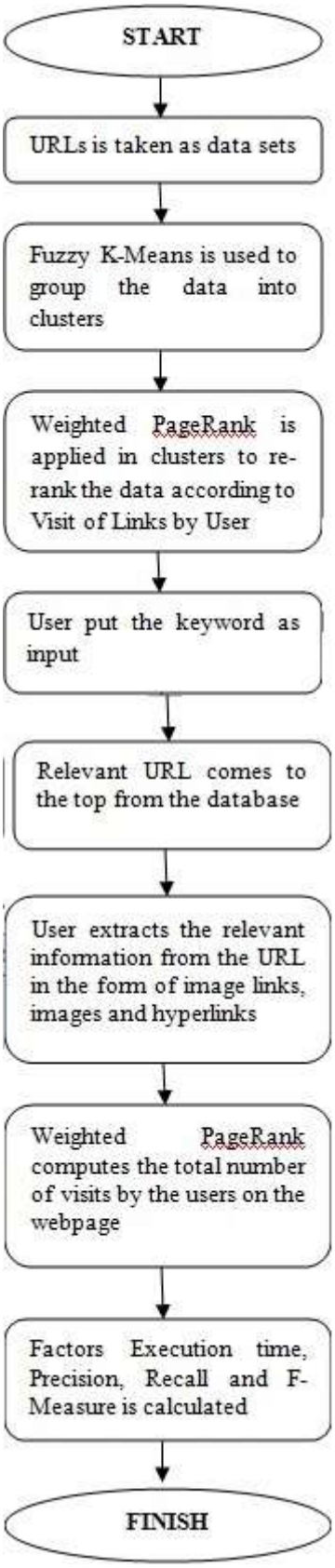

Figure 1: Methodology of Proposed Work

5. Results

We have run our proposed approach on various top most visited URLs such as tanishq, yahoo, titan, ebay, homeshop18, twitter etc. An Example as output is as shown

below. In this project hybrid technique is used to extract the dynamically image links, images and hyperlinks from the URL. Execution time is also fetched to extract the hyperlinks from URL.

Figure 2: User put the Keyword as Input

Figure 3: URL is retrieved

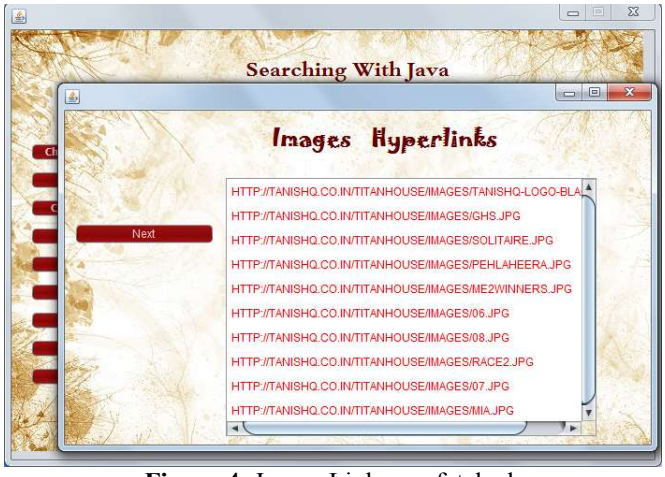

Figure 4: Image Links are fetched

Figure 5: See the Images in Image Panel

Figure 6: Total Hyperlinks

Figure 7: Execution Time to fetch the Links

Figure 8: Weighted PageRank

Figure 8: URL Availability in another URL

5.1 Parameters Analysis

At the end of project, we have experiment with hybrid approach and compute various parameters to run the top most visited URLs. We have done analysis on valid measures such as Execution Time, Precision, Recall, F-measure and G-measure.

Execution Time: we define the execution time in our project as to time taken to retrieve the image links and hyperlinks from the URL. The unit of time taken is milliseconds (ms).

Precision: The fraction of retrieved documents that is relevant to find:

$$Precision = \frac{Total\ Relevant\ documents}{Total\ Retrieved\ documents}$$

Recall: The fraction of the documents that is relevant to the query that is successfully retrieved:

$$Recall = \frac{Retrieved\ Relevant\ documents}{Total\ Relevant\ documents}$$

F-Measure: The harmonic mean of combination of precision and recall:

$$F - Measure = \frac{2 * Precision * Recall}{(Precision + Recall)}$$

G-Measure: The geometric mean of combination of precision and recall:

$$G - Measure = \sqrt[2]{Precision \times Recall}$$

Table 1: Results of Proposed Method

ID	*URLs*	*Time Taken(ms)*	*Precision*	*Recall*	*F-Measure*	*G-Measure*
1.	http://www.naturewallpaper.net/	5782	0.4285	0.8888	0.5783	0.6172
2.	http://www.easychair.org/	4324	0.75	0.3333	0.4615	0.5
3.	http://titan.co.in/watches	5497	0.6631	0.1005	0.1745	0.2581
4.	http://all-free-download.com/free-photos/rose-flowers.html	8182	0.2086	0.8333	0.3337	0.4170
5.	http://www.ebay.in/	6050	0.0902	0.9459	0.1647	0.2921
6.	http://www.homeshop18.com/	5295	0.3555	0.75	0.4824	0.5163
7.	http://adaptive-images.com/	5271	0.6071	0.7647	0.6767	0.6813
8.	https://twitter.com/	8705	0.0961	0.2	0.1298	0.1386
9.	https://in.yahoo.com/?p=us	8919	0.6847	0.9365	0.7911	0.8008
10.	http://jewellery.picturesklix.com/	5666	0.3596	0.9024	0.5143	0.5697
11.	http://tanishq.co.in/Home	4065	0.1458	0.7142	0.2422	0.3227

6. Conclusions

In this paper, we have proposed a hybrid technique of clustering and ranking mechanisms such as Fuzzy K-Means and Weighted PageRank based on Visit of Links which are used to fetch the images, image links and hyperlinks from the URL. In an existing work, Weighted PageRank and hybrid approach such as K-Means with Weighted Pagerank or PageRank has done on text or URL to retrieve the text and contents but has not done on links. But this paper proposed hybrid approach is performed on links and images to provide the relevant result to the user. And we have done experiment with our approach and compute the various parameters with good results. In the background history, we analyzed Fuzzy K-Means has several advantages over K-Means and other clustering technique and same as Weighted PageRank based on Visit of links has its own strength over previous approaches such as PageRank, PageRank based on Visit of Links and orginal Weighted PageRank. Now, we combined both this techniques and provide good results in this paper. So, we can say that our hybrid technique can see as future of web mining.

7. Future Scope

The web mining trend is increasingly very rapidly. Nowadays trillion of users fetch the information from web. But to retrieve the relevant and required information, Web mining techniques play an important role, similarly our paper described a hybrid technique by combining well known two algorithms Weighted PageRank based on Visit of Links and Fuzzy K-Means which is used to extract the rich information in the form of relevant links and images. Today, users

download huge number of images from the search engine. But sometime User does not get relevant images or links. But our technique provides very fast quality and relevant search result with user. So that we can say our approach can be efficiently use by search engine to retrieve the required links and images to provide with user. And hence, our approach can be seen as future of web mining.

References

[1] S. Brin, and Page L., "The Anatomy of a Large Scale Hypertextual Web Search Engine", Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.
[2] Wenpu Xing and Ghorbani Ali, "Weighted PageRank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.
[3] Miguel Gomes da Costa Júnior and Zhiguo Gong, "Web Structure Mining: An Introduction", In Proceedings of the 2005 IEEE International Conference on Information Acquisition, July 3, 2005
[4] Gyanendra Kumar, Neelam Duahn, and Sharma A. K., "Page Ranking Based on Number of Visits of Web Pages", International Conference on Computer & Communication Technology (ICCCT)-2011, 978-1-4577-1385-9.
[5] G. Shrivastava, K. Sharma, V. Kumar," Web Mining: Today and Tomorrow", In Proceedings of 2011 3rd International Conference on Electronics Computer Technology (ICECT), pp.399-403, April 2011
[6] Syed thousif hussain, B. N.Kanya and Dr.MGR, "EXTRACTING IMAGES FROM THE WEB USING DATA MINING TECHNIQUE", International Journal of Advanced Technology & Engineering Research, VOLUME 2, ISSUE 2, MARCH 2012
[7] Neelam Tyagi and Simple Sharma, "Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012
[8] Preeti Chopra and Md. Ataullah, "A Survey on Improving the Efficiency of Different Web Structure Mining Algorithms", International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume-2, Issue-3, February 2013
[9] Monika Yadav and Mr. Pradeep Mittal, "Web Mining: An Introduction", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013
[10] Gurpreet Kaur and Shruti Aggarwal, "Improving the Efficiency of Weighted Page Content Rank Algorithm using Clustering Method", International Journal of Computer Science & Communication Networks,Vol 3(4),231-239.
[11] Gurjit Singh and Navjot Kaur, " Hybrid Clustering Algorithm with Modifications Enhanced K-Means and Hierarchal Clustering", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3,Issue 5, May 2013.
[12] Seifedine Kadry and Ali Kalakech, "On the Improvement of Weighted Page Content Rank", Journal of Advances in Computer Networks, Vol. 1, No. 2, June 2013
[13] Rashmi Rani and Vinod jain, " Weighted Page Rank using the Rank Improvement", International journal of Advanced Research in Computer Science and Software Engineering, volume 3,Issue 7, July 2013.
[14] Supreet Kaur and Usvir Kaur, "An Optimizing Technique for Weighted Page Rank with K-Means Clustering", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013.
[15] Jaideep Srivastava, Prasanna Desikan and Vipin Kumar, Web Mining— Concepts, Applications, and Research Directions
[16] Precision, Recall and F-measure Definition, Available: http://en.wikipedia.org/wiki/Precision_and_recall
[17] Web Mining Definition, Available: http://en.wikipedia.org/wiki/Web_mining

Author Profile

Rashmi Sharma received the B.Tech. degree in Information Technology from Punjab Technical University, Jalandhar, Punjab, India in 2012. Currently, she is pursuing M. Tech in Computer Science and Engineering at Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India.

Kamaljit Kaur received the B.Tech. and M.Tech degree in Computer Science and Engineering. She is pursuing PhD in Data and Web Mining. Currently, she is an Assistant Professor at Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India. She has over 10 year's job experience and she is guiding various M.Tech. Thesis in the field of Database Security and Data Mining.