International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), Nov 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

IMPROVED WEB INFORMATION RETRIEVAL USING CONTENT'S WEIGHT BASED PAGE RANKING AND DYNAMIC CLUSTERING

Charanjit Singh[1], Vijay Laxmi[2], Arvinder Singh Kang[3]

[1]*Research Scholar, Guru Kashi University, Talwandi Sabo(India)*

[2]*Professor and Dean, Guru Kashi Universit, Talwandi Sabo(India)*

[3]*Professor and Dean, Chandigarh University, Gharaun(India)*

ABSTRACT

Today's search engines are retrieving thousands of web pages for a single query, where most of the results are irrelevant as per the query entered by the user. Therefore, there is very real necessity to listing the results according to the user needs. The big challenge lies to ordering retrieved pages, organizing them and presenting them to users in line with their interests. The most of search engines utilizes the page ranking algorithms to analyze web pages and re-rank search results according to the relevance of the user's query by calculating the importance of a web page. The proposed work examines different web page ranking algorithms/methods and their recently-developed improvements. In addition to proposed framework, a new ranking technique called Content's Weight based Page Rank is developed for implementation, whereas for organizing, hierarchical clustering uses dynamically and finally presenting results based on proposed ranking algorithm as per the user's interest. An experimental setup is defined over proposed framework called Dynamic Searching System to get performance measure metrics i.e. Precision, Recall and F-measure. The results validate the effectiveness of the proposed work by measuring performance metrics and comparing with existing work demonstrate its efficiency.

Keywords – Web mining, World Wide Web, Search Engine, Web Page, Page Ranking, Clustering.

I. INTRODUCTION

To retrieve information [1] from the web resources, World wide web (WWW) plays a crucial role. A tool called search engine is used to retrieve the required information from the web matrix. In general search engine, it crawls the web page's content from the various nodes and organize them in list of resultant pages to the user so that they can easily access the required information from the web pages by their provided links. In earlier as per the user's request, this approach implemented well because their resources are limited. Users were well capable to recognize the relevant information from the search engine results. By increasing in usage of internet, the concept of resource sharing is also increases. This leads to adopt an approach where ranks should be assigned to each web content resources automatically.

Web Mining

Data mining, text mining, web retrieval and information retrieval [2] are the research areas which are more crucial to extract data from WWW. Whereas web mining is the research area which concluded all above said

research areas. Web mining can be classifying on two basic aspects i.e. the purpose and the data sources. Retrieving relevant data from the existing data or large database of documents repository is the main focus of Retrieval, whereas the mining research is mainly focus on discovering new information from the data. Therefore, the Web mining can be classified into:

- Web structure mining
- Web content mining
- Web usage mining

Web structure mining [3] is used to generate the structural summary of the website and webpage with respect to extract the patterns from hyperlinks of web. The structural component of web page is hyperlink which study the connection of web pages to different location.

Web content mining is used to extract useful information from the content of the web page [5] with respect to collection of facts in web page. Content mining is related to data and text mining because various techniques for the same can be applied and web content are text based in it. Whereas different due to semi-structured and or unstructured data.

Web usage mining [6] is also an application of data mining techniques [7], used to discover usage patterns from Web data tends to understand and better serve needs of Web based applications. Three major phases consist i.e. pre-processing, pattern discovery and pattern analysis.

As per the increasing in web resources and competition, the ranking of web become monotonous and dynamic in nature as per the query of users. Various ranking criteria used by search engines to rank the web resources for the query of user. This tends to business motivation of taking up their web resource onto to the high-ranking position of web resource. Different ranking algorithms [8] considered to rank the web pages as per the specification. Certain ranking approaches are:

PageRank Algorithm

Page ranking [6] is most commonly used approach to rank the web page and measure the importance of it. According to this algorithm, rank of the page defines and depend on the number of all incoming link to it. On the same time, outgoing links of the page also become important compare to incoming links. A page receives high rank itself, if a page is linked to many pages with high page rank. Several iterations require [9] to be executed by the page rank algorithm and after each iteration, values will be approximated better to real value. The following expression 1 used at each iteration for each web page.

$$PR(u) = d \sum_{v \in Bu} \frac{PR(v)}{L_v} \quad (1)$$

Here, ‘d’ is a factor used for normalization, ‘u’ as a web page, B_u as the set of pages linked to ‘u’, PR(u) and PR(v) are rank scores of pages’ ‘u’ and ‘v’, respectively, and L_v denotes outgoing links of page ‘v’. The final page ranking algorithm formula is as given bellow:

$$PR(u) = (1-d) + d \sum_{v \in Bu} \frac{PR(v)}{L_v} \quad (2)$$

Here, 'd' is a damping factor and it usually set to 0.85. Basically, 'd' can be as the probability of users that following direct link, (1-d) denotes as the page rank distribution from pages that are non- directly linked.

Weighted Page Rank Algorithm

It's an extension of page rank and use to assign rank [6] according to their importance or popularity compare to page rank dividing it evenly. Popularity assigned in term of weight values to in-link denoted as $W^{in}_{(v,u)}$ and out-link $W^{out}_{(v,u)}$ respectively. W^{in} denotes as the weight of link (v, u) that calculated based of incoming links to page 'u' and also no. of links (incoming) to all outgoing links pages of page 'v', as shown in following expression as:

$$W^{in}_{(v,u)} = I_u / \sum_{P \in R(v)} I_P \quad (3)$$

Here, I_u and I_p shows the no. of incoming links of page 'u' and 'p' respectively. R(v) as the reference page list of pages 'v'. W^{out} shows as the weight of link(v,u) that is calculated based on no. of outgoing links of page 'u' and no. of outgoing links of reference pages of page 'v', show in equation 4.

$$W^{out}_{(v,u)} = O_u / \sum_{P \in R(v)} O_P \quad (4)$$

Here, outgoing links of page 'u' and 'p' is represented by O_u and O_p respectively. Then final weighted page rank equation 5 is as follow:

$$WPR(u) = (1-d) + d \sum_{V \in B(u)} WPR(v) W^{in}_{(v,u)} W^{out}_{(v,u)} \quad (5)$$

Page Content Rank

A new ranking method defined called page content rank [10] that is based on page relevance. According to this approach, to seem the importance of page is analyzed on the content of web page. The importance of page is based on the importance of terms present in web page, while to specify the importance of term is based on given query 'q'. Therefore, this approach uses the neural network and structure of its inner classification.

The calculation of term 't' importance denoted by importance(t) and carried out basis of 5=(2*NEIB) parameters, where NEIB state as no. neighboring terms that is included in the calculation. Database 'D', query 'q' and the number n of pages are attributes on which calculation depends. Furthermore, the classification function called classify () used with 5+(2*NEIB) parameters returns the importance of 't'. The importance of term 't' considered by certain parameters such as Term extraction, Term Classification, Relevancy Calculation and Term Frequency. In this research work, Term frequency is considered as per the following expression:

$$freq(t) = \sum_{P \in R_q} TF(P,t) \tag{6}$$

The said expression helps to determine the total number of occurrence of defined term 't' in R_q. This also become interest of users to choose the search engines to identify the relevant information as per their needs. So, there is requirement to develop a novel approach to ranking the web resources as per their contents based on the query of user.

II. PROPOSED FRAMEWORK OF DYNAMIC SEARCHING SYSTEM

A framework is developed called Dynamic Searching System, web pages are fetched which are relevant to user's query. There are two major components are in this system as ranking web pages and clustering web pages dynamically. These two components are back bones of the system, which helps together to get the relevant pages from the web page database. Dynamic Searching System an enhanced searching system, it helps the users to search relevant web pages from its defined clusters which are stored in cluster database. In this system, if query keyword has its cluster then it searches from cluster and send back to the user according to proposed ranking approach. The new cluster created dynamically if new keyword entered in the system as a user's query. The interaction of the DSS between the users and process of ranking with their cluster database to fetch the relevant web page from the web page database with their explanation as given in Fig. 1.

The details of various components of DSS are as:

- **Searching System Interface:** it is the front end of the searching system and used to enter a keyword as a query and show results as output.
- **Web page Repository:** there is a database of web pages which stores the various URL links with their keyword identification on arrival on a new keyword. There is a crawler, helps to crawl web pages from word wide web for the particular keyword and store in web page repository. The database provides the list of URLs which are relevant or associated with that new keyword whenever users enter a new keyword.
- **Ranking Module:** This module of system, assign rank to each URLs. When user enters a keyword as a query that associated with one of cluster and already stored in the cluster database. This module accepts the query, fetch the associated URLs from cluster and apply proposed ranking algorithm i.e. CWPRA (described in next section).
- **Cluster Database:** The sets of various web pages are stored in different clusters with respect to their domains in the database of cluster. Proposed framework has certain domains with their clusters, keywords and associative URLs in cluster database initially. When user enters a keyword as a query, the system searches in different domains of database. The database should be consistently increasing also because if a new keyword is being searched in system.
- **Dynamic Clustering**: If a new keyword as a user's query entered in system, there is no any appropriate URLs in any cluster of database. The concept of dynamic clustering in system helps to creating a new

cluster on arrival of new keyword dynamically (as per requirement) in database under its respective domain for that new keyword.

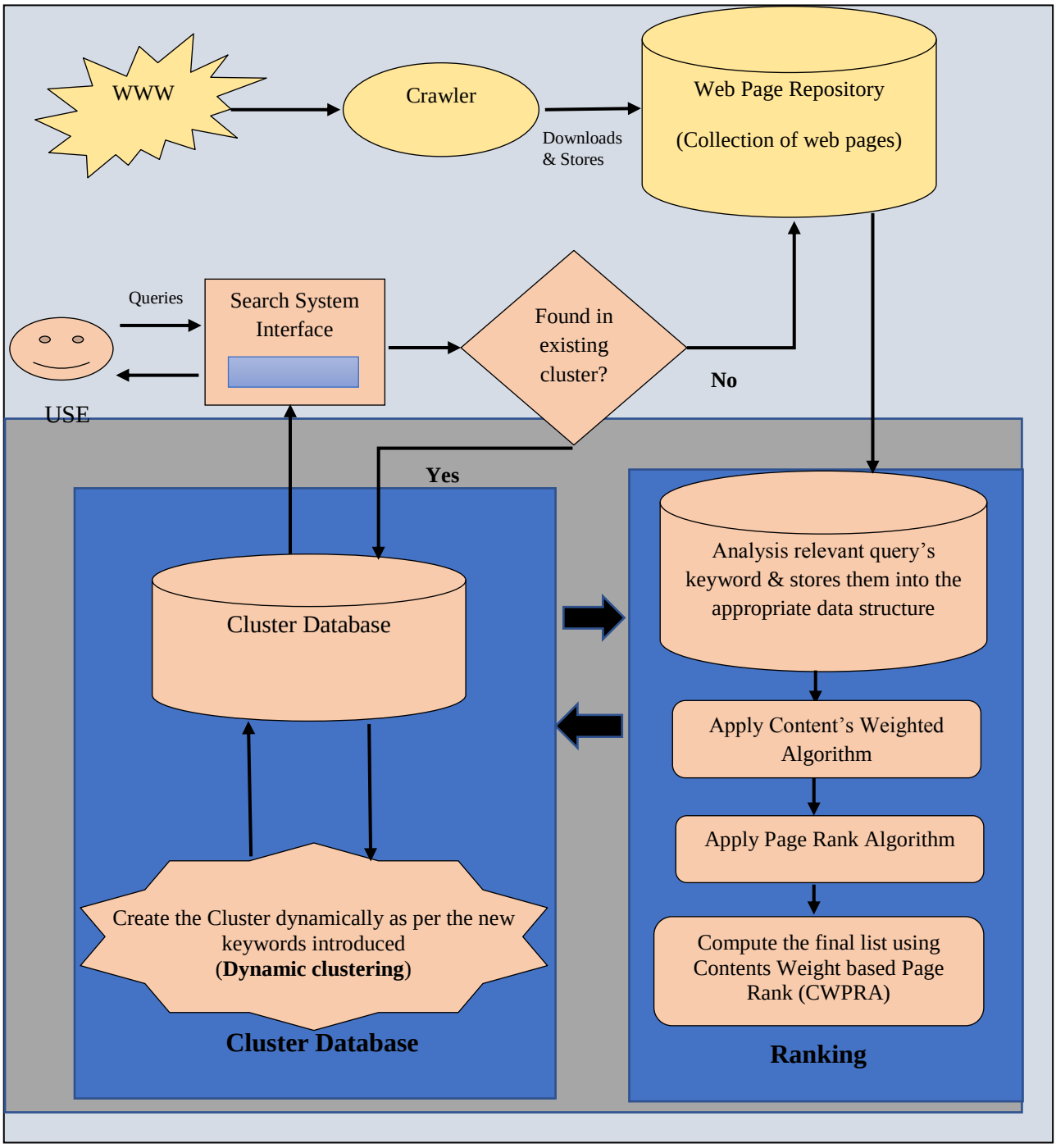

Fig1. Framework for Dynamic Searching System

The concept of data pre-processing plays an important role for knowledge based system. Basically, to improve the accuracy and relevancy and to improve the quality of the data the pre-processing makes an efficient move. In

proposed framework architecture, the concept of pre-processing become a crucial phase while using text mining due to incomplete and inconsistent real-world data. The local database is used to store all keywords with respect to their URLs. The system accepts the keyword as query from the system's interface and after matching the keyword in local database, their all respective URLS has been fetched. A new list of fetched URLs has been constructed against the keyword after extracting keywords from each link of URL. The pre-processing consists of all these tasks i.e. processed request, extracted URLs for the keyword and finally stored in the specific data structure. After this pre-processing task, the list of URLs passes over the ranking part of the system to assign the rank to each URL as defined in the proposed ranking approach.

For ranking the URL, the proposed ranking approach is use to assigning the ranks each of the URL of list. The process of ranking has been defined in three major steps: firstly, count frequency of keyword from of the URL, second compute content weight of URL based on frequency of keyword and in-links of URL and finally compute content weight page rank using content's weight and all links of URL. The process of pre-processing and ranking the URLs as per the architecture shown in Fig - 2.

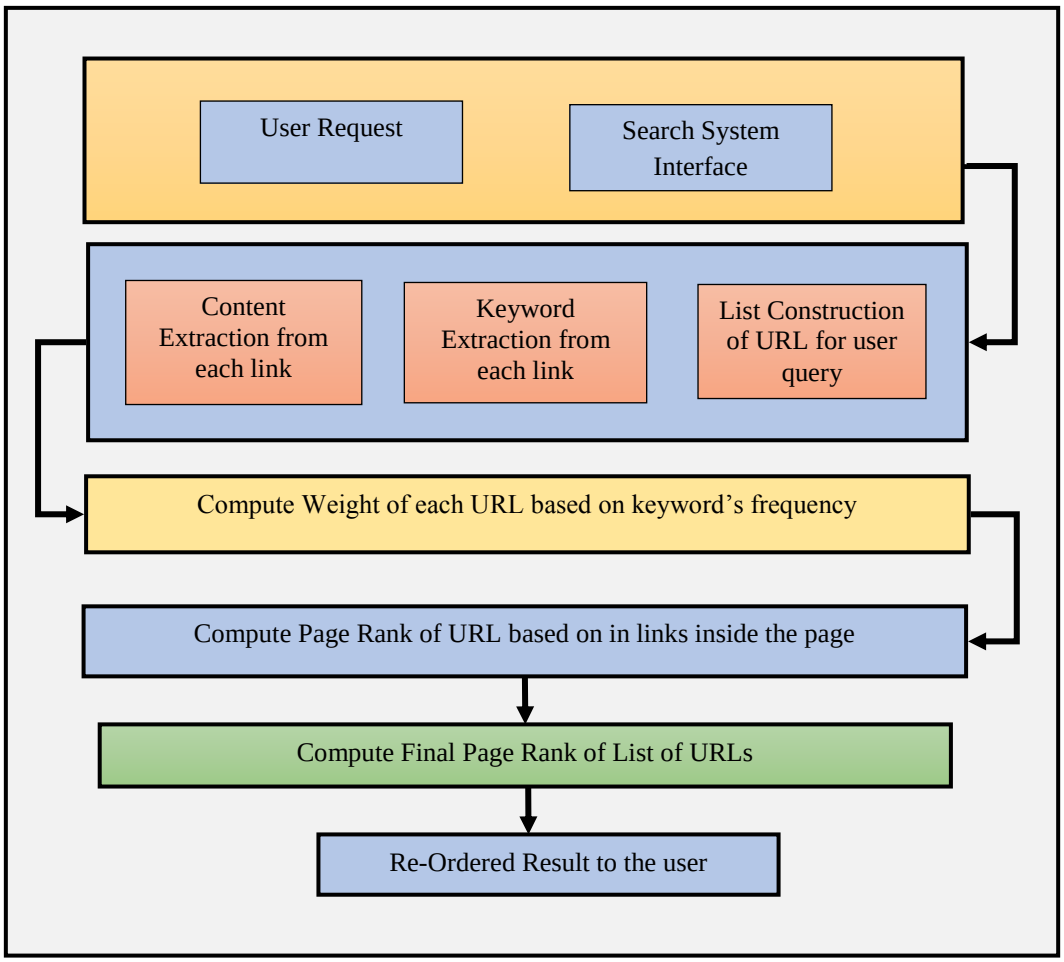

Fig2. Computation of CWPRA

The content's weight based page ranking algorithm is used in proposed system to rank the list of URLs. The CWPR algorithm is implemented and imbedded in proposed framework.

The following steps of CWPR Algorithm followed by proposed system to assigning rank to list of URLs:

(Pre-processing)

1. User enters the keyword (N) as the input in system.
2. If keyword is found, then respective URL's links has been fetched from database and stored in array (*X*). i.e. Array(X) = {k=0 and nn=X.length, where *X*= $link_1$, $link_2$, $link_3$ ……., $link_{nn}$}.
3. Now algorithm apply the following task on URL

a) Remove stop words

b) Apply stemming on each line of content in page

c) Concatenate to String variable (Q_n)

4. After getting the concatenated string variable (Q_n), find the string length (Z) i.e. *Z= strlen (Q_n).*

(Ranking Part)

5. Now, from 0 to till Z, algorithm will count the frequency of keyword (X_r) from each URL.
6. After that, total weight (T_w) is calculated with respect to all query parameters, frequency of keyword and length of string as per the no. of keywords in the URL. i.e. $(T_w)= \frac{X_r+D}{Z}$.
7. After that calculates the total no. of links (T_l) by finding the all the in-links (I_l), out-links (O_l) and External Image links (EI_l) from URL. i.e. $T_l = I_l + O_l + EI_l$
8. Now algorithm finds Content's weight for the URL as with respect to weight and in-links. CW= $I_l + T_w$ (*for example content's weight for web A will be CW(A)=* $I_l + T_w$).
9. Further, algorithm calculates the final Content's Weight based Page Rank (CWPRA) of a URL (*i.e. A*) by using basic page ranking criteria, where add all web page's relative content's weight over total no. of links exists in it. i.e. CWPR(*A*)= $\sum \frac{CW(A)}{T_l}$.
10. After getting the all ranks of URLs, an associative array AA is used to store the ranks. Now, array will be arranged in descending order as per their ranks and send to the user as output.

III. EXPERIMENTAL SETUP

The experimental setup for proposed Dynamic Searching System using an archetypal model LAMP. This model has four open source components i.e. Linux operating system, the Apache HTTP server, the MySQL as relational database management system and finally the programming language PHP. The LAMP architecture usually used for dynamic web sited and other interactive web services. To handle the transaction between front-end, its respective back-end and associated data base in this framework an Apache HTTP server is used. For

back-end process, PHP server-side language has used. To edit pull and edit the information in the database, the PHP has been designed. This PHP is normally collaborated with the databases which are written in SQL language. Therefore, in this framework to deal with clusters database MySQL has used.

Two main components called CWPR algorithm and Hierarchical clustering are the backbone behind the proposed system. A developed searching system finds the set of relevant web pages from the database. If the inserted keyword already has its respective cluster then system find that cluster and ranked their URLs as per the ranking algorithm and show to the user as the result. On the other hand, if new keyword is introduced then system search its respective URLs from the WWW, grouped as a cluster and finally ranked them and show to the user as the results.

Dataset

The database for the system as various sets of URLs with its respective keyword are collected in June 2017 and grouped as cluster with respect to its domains. Initially it contains 9 different domains having more than 400 Urls that are stored in different clusters. As per the proposed system, database of the system has been growing automatically as user introduced new keyword. The following 'Table No 5 Cluster Database' shows the organization of the cluster database where the keywords and its URLs are organized:

Table1: Cluster Database

Sr. No	Domains	Clusters
1	Technology	Laptop, mobile, smart watch, tracking device, pen drive, network
2	Sports	cricket, hockey, football, tennis, swimming
3	Education	Science, computer, agriculture, electronics
4	Fashion	Jewellery, ornament, jeans, shirts, beauty
5	Animals	Tiger, jaguar, horse
6	IT	Data mining, PHP, Web, Google
7	Eatables	Apple, mango, kiwi, carrot, radish
8	Furniture	Chairs, beds, tables, racks, bookcase, sofa
9	medicines	Fever, pain, cancer, infection, allergy, nutrition

Performance Metrics

The performance of proposed searching system is compared with existing searching method using performance measures like precision, recall and f-measure. These factors are computed as below:

1) *Precision:* Precision is the fraction of relevant URLS retrieved and the total number of URLs retrieved by the system.

$$Precision = \frac{relevant\ document \cap retrieved\ documents}{retrieved\ documents} \quad (6)$$

2) *Recall:* Recall is the fraction of documents retrieved successfully and relevant to a query. It's also called sensitivity, can be looked at as the probability that a relevant document is retrieved by the query.

$$Recall = \frac{relevant\ document \cap retrieved\ document}{relevant\ documents} \quad (7)$$

3) *F-measure:* A harmonic mean of precision and recall is f-measure. It also provides good results when there is a good result of precision and recall itself.

$$F - measure = 2.\frac{precision * recall}{precision + recall} \quad (8)$$

IV. RESULTS

Experiments are conducted using different queries and 10 results are ranked and analyzed for each query and to check the performance of the retrieved results based on the metrics like precision, recall, and f-measure and are shown from Table 2 to Table 4 respectively. As per user interest the value of precision varies. The given values of precision values define the relevancy of search results obtained during experimentation. Search recall values, which are measurement of search accuracy are also searched in this section. F-measure is also calculated here by considering the estimated precision and recall values and the results for the same are listed below:

Table 2. Comparison of Precision Values of Existing and Proposed Ranking Approach

Queries	**Precision Values for different Ranking Algorithm**				
	PR	**WPR**	**HITS**	**Existing Page Ranking**	**Proposed Page Ranking**
Apple	0.862	0.904	0.961	0.991	0.949
Data mining	0.886	0.870	0.952	0.926	0.948
PHP	0.799	0.893	0.904	0.919	0.947
Web	0.589	0.791	0.842	0.993	0.950
Jaguar	0.647	0.751	0.835	0.893	0.955
Google	0.798	0.812	0.847	0.885	0.948
Network	0.719	0.729	0.787	0.945	0.949
Tiger	0.731	0.771	0.753	0.932	0.959
Average	**0.754**	**0.815**	**0.860**	**0.936**	**0.951**

Table 3. Comparison of Recall Values of Existing and Proposed Ranking Approach

Queries	Recall Values for different Ranking Algorithm				
	PR	**WPR**	**HITS**	**Existing Page Ranking**	**Proposed Page Ranking**
Apple	0.977	0.979	0.970	0.989	0.990
Data mining	0.993	0.937	0.950	0.960	0.989
PHP	0.968	0.943	0.971	0.979	0.990
Web	0.974	0.987	0.958	0.980	0.992
Jaguar	0.978	0.876	0.984	0.990	0.992
Google	0.941	0.969	0.940	1.000	0.991
Network	0.969	0.947	0.947	1.000	0.991
Tiger	0.945	0.940	0.945	0.952	0.991
Average	**0.968**	**0.947**	**0.958**	**0.981**	**0.991**

Table 4. Comparison of F-Measure Values of Existing and Proposed Ranking Approach

Queries	F-Measure Values for different Ranking Algorithm				
	PR	**WPR**	**HITS**	**Existing Page Ranking**	**Proposed Page Ranking**
Apple	0.380	0.917	0.962	0.990	0.971
Data mining	0.926	0.910	0.951	0.943	0.962
PHP	0.869	0.864	0.895	0.947	0.971
Web	0.847	0.737	0.895	0.986	0.975
Jaguar	0.849	0.743	0.902	0.938	0.979
Google	0.872	0.875	0.890	0.938	0.974
Network	0.831	0.816	0.859	0.991	0.979
Tiger	0.848	0.822	0.837	0.941	0.965
Average	**0.803**	**0.836**	**0.899**	**0.959**	**0.972**

As per the obtained results in table 2, there is comparison of proposed ranking approach to existing ranking approaches therefor the performance of the proposed technique is optimal. Using developed system, recall values shows in table 3 are very high approximately retrieving all documents in response to a query. Table 4 shows that the proposed ranking approach provides good results because by using precision and recall values computed f-measure values are better as compared to existing ranking techniques.

So, finally we concluded the comparison of average values of Precision, Recall and F-Measure for various ranking algorithm's in Table 5 and graphically represents the comparison in fig 3. From this table, this is concluded that the proposed system produces better values from existing one.

Table 5. Comparison of Average Values of Precision, Recall and F-Measure

Parameters Vs. Algorithms	**PR**	**WPR**	**HITS**	**Exiting Algorithm**	**Proposed Algorithm**
Precision	0.754	0.815	0.860	0.936	0.951
Recall	0.968	0.947	0.958	0.981	0.991
F-Measure	0.803	0.836	0.899	0.959	0.972

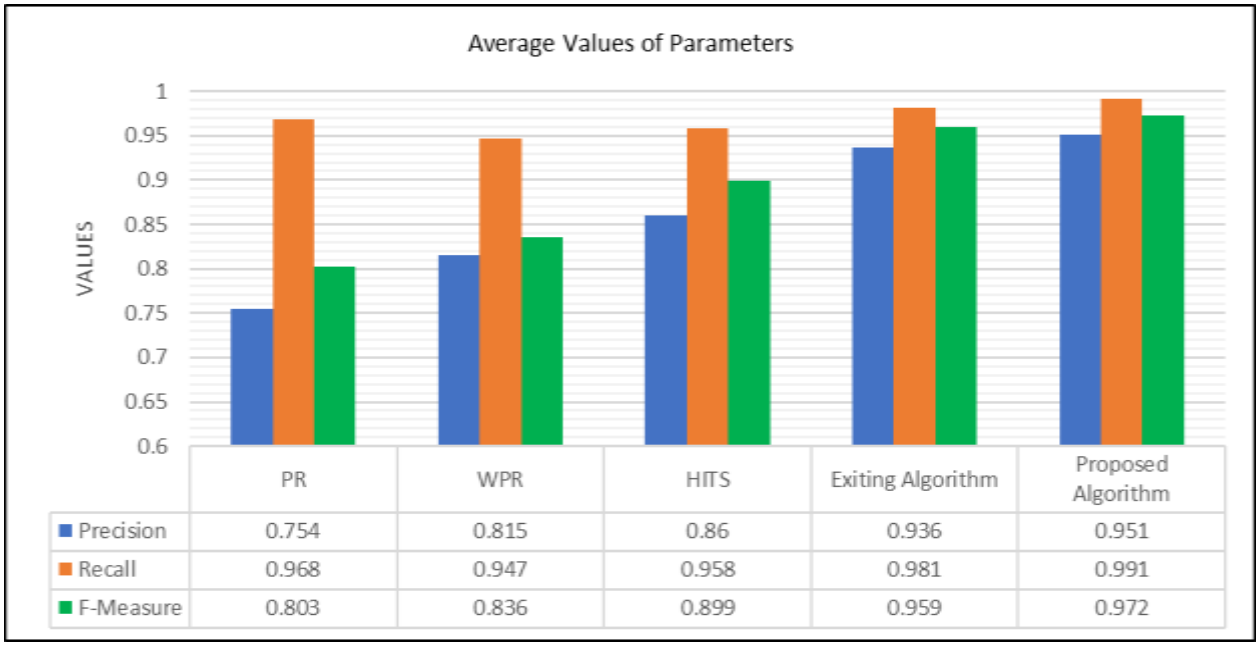

Fig3. Graph Showing Comparison of Average values of Parameters

V. CONCLUSION AND FUTURE SCOPE

The proposed framework called Dynamic Searching System is intended to provide relevant web pages as the result as per user's query. In this system, a new ranking approach called Content's Weight based Page Ranking is developed and provides rank as per the content's weight and their links of the web pages. In make improvement to retrieving most relevant web pages by system, a hierarchical clustering technique is used. In this system, on initializing a new keyword, clusters of different domain are created using the concept of dynamic clustering. To measure the relevancy of searching system, various measuring parameters are evaluated i.e. Precision, Recall and F-measure. The improved results from the existing one demonstrates the high efficacy of relevancy of search results.

In future, the proposed system can be validated on searching the relevant web pages on the basis user's interest or behaviour of the users using web usage mining.

VI. ACKNOWLEDGEMENT

I would like to express my deep gratitude and respect to Dr. Vijay Laxmi whose gives endless help me in this research. Further I also thanks to Dr. Arvinder Singh Kang whose advices and insight was valuable to me for my research work and make this happened. I highly thankful for all I learned from them. I would also be thankful for encouraging and helping me for my research work on all the stages and being an open person to idea. Their researching skills always support me to shape my ideas and make this a better research for the society.

REFERENCES

[1] P. Sudhakar, G. Poonkuzhali and R.K. Kumar, "Content Based Ranking for Search Engines," International Multiconference of Engineers and Computer Scientists (IMECS '12), Hong Kong, 2012.

[2] P. Sharma, D. Tyagi and P. Bhadana, "Weighted Page Content Rank for Ordering Web Search Result," *International Journal of Engineering Science and Technology (IJEST), 2(12) 7301 – 7310, 2010*.

[3] S. Chakrabarti, B.E. Dom, S.R. Kumar, P. Raghavan, et al., "Mining the Web's link structure," *Computer, 32(8) 60 – 67, 1999.*

[4] W. Jicheng, H. Yuan, W. Gangshan, and Z. Fuyan, "Web mining: knowledge discovery on the Web," IEEE International Conference on Systems, Man, and Cybernetics, Tokyo, Japan, 1999.

[5] R. Kosala and H. Blockeel, "Web Mining Research: A Survey," ACM SIGKDD Explorations Newsletter, New York, USA, 2(1), 2000.

[6] J. Srivastava, R. Cooley, M. Deshpande and P. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," ACM SIGKDD Explorations Newsletter, New York, USA, 1(2), 2000.

[7] A.A. Barfourosh, H.R. Motahary Nezhad, M.L. Anderson and D. Perlis, "Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition," Technical Report, University of Maryland, 2002.

[8] O. Etzioni, "The World Wide Web: Quagmire or gold mine," Communications of the ACM, 39(11) 65-68, 1996.

[9] J. Pokorny and J. Smizansky, "Page Content Rank: An Approach to the Web Content Mining," IADIS International Conference on Applied Computing, Algarve, Portugal, 2005.

[10] Singh, C., and Kautish, S.K., (2015), "Page Ranking Algorithms for Web Mining: A Review", International Conference on Advancements in Engineering and Technology, Sangrur, India, 20-21 March 2015, pp. 406-410.

[11] Dhiliphan, R.T., Suruliandi, A., and Selvaperumal, P., (2015), "Content and user click based page ranking for improved web information retrieval", *International Journal on Computational Science & Applications, 5(6), pp. 111-127.*

[12] Singh, C., Laxmi, V., and Kang, A.S. (2017), "Dynamic Clustering Case Study using K- Mean Clustering Algorithm" *International Journal of Computer Science and Information Technology & Security, 7(4), pp. 19-22.*