Ranking Web page based on Content's Weights for Search Engines

Charanjit Singh
Research Scholar
Guru Kashi University, TalwandiSabo,
Bathinda, India
sehgal_cs@yahoo.com

Vijay Laxmi
Guru Kashi University, TalwandiSabo, India
deanca.gku@gmail.com

Arvinder Singh Kang
Chandigarh University, Gharaun, India
dean.sw@cumail.in

Abstract- **In today's era where everything is on the webpages, to searching, retrieving and organizing information around the world, search engines plays an important role. Although in actual position, the relevance of results by search engines as per the user's query stills a big challenge for the researchers. The search engines return the vast amount of unexpected and repeated results. Various ranking criteria used by search engines to order the results. Web mining and ranking are the most powerful research areas, use to enhance process of retrieving the relevant information in better and faster manner from the web. Web content mining and web structure mining, concepts of web mining play a major role in search engines. Web content mining is used to deal with content of the web page whereas structure mining with links of web pages. In this research work, enhanced ranking approach is proposed called Content's weight based page ranking algorithm. In this ranking algorithm, rank has been assigned to the web page on the basis of their contents, its respective weight and page rank with in-links. A search engine environment is developed using proposed ranking algorithm to retrieve the relevant web pages as per the user's query. Experimental mathematical calculation results show that how the sequence of dataset has been changed or improved on the basis of their new ranks computed using the proposed ranking manner.**

Keywords: Content's Weight-based Web Retrieval, Web based Services, Search Engines, Ranking Web Page.

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.

A. 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:

(a) Web structure mining

(b) Web content mining

(c) 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:

1) 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 is 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.

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

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:

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

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.

2) 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:

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

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.

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

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:

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

3) 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. neighbouring 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 RELATED WORK

To develop automatic retrieval mechanism, numerous challenges faced by researchers such as lack in web data structure, heterogeneity of network resources etc. Furthermore, the nature of web data as unstructured and semi- structured implies the need of web content mining where the ranking is not just based on in-links and out-links of web page. Author defines [11], the two different points of view such as information retrieval and database view of web content mining. Research [12], defined various issues of content mining and various characteristics of web. Web mining's research area and various different categories briefly discussed in research paper [13]. Also, their research work concluded semi-structured and unstructured data from IR. In information retrieval (IR) view, semi-structured represented by hyperlink structure and HTML structure whereas unstructured text as bag of words.

In paper [3], a method defined to assign relevance ranking to the web pages according to the user's query. They also stated that to transform a web site into database, the mining always attempts to infer the structure of the web site in database view (DB). Numerous problems occurred to recognizing content for example sequence labelling problem stated as the common structure problem in natural language processing and machine learning processes is acknowledged in [15]. Useful knowledge become presentable by the extracted structured and semi structured data with mining process using a web content mining survey as a tool in [16]. A framework is proposed [17], to facilitate a searching. In framework user get the information about any product without visiting the homepage of the companies in spite using Query interface user enter the product name and project searches web pages available related to the text entered.

Retrieving relevant information from structure and unstructured documents using Statistical approach of proportional and chi-square in [10] [18]. They applied method to detect and remove redundant web document using correlation method. Paper stated that today's users rely on web search engines to retrieve information such as Yahoo, Google or Bing for specific information, a huge amount of results are returned both relevant or irrelevant. Therefore, to discovering the vital information from resources of web pages, web mining research community has to take essential steps. The same name of ranking algorithm [4] which rank the webpages on the basis of content and weight. In this paper, they defined the procedure with limited algorithm specification.

An approach proposed [1], in which dictionary for the extracted web is created by pre-processing process, process the relevancy and finally rank the web pages on the basis of keyword and content from the web pages. The criteria defined [2], in which web page is being rank as per the relevancy extracted and compute the weight for the same and get the final rank of the web page. They also defined the difference of the various ranking methods. A crawler focused system defined [3], called a hypertext resource discovery system. This crawler analyses its boundary, the most relevant for the crawler are lies in its boundary and also avoids irrelevant region. A new approach defined [19], to resolved certain problems such as noise, slow retrieval speed and links that is broken on the basis of retrieval information and hazards on needs of future.

The concept of indexing [20] explored, using HAIRCUT (Hopkinks Automated Information Retrieving for Combing Unstructured Text) and N-grams system. A method presented [21], that is efficient and used to finding duplicate copy of document collection that is used to improve crawler of web, ranking functions and archivers in search engines. A method proposed [22], that is information retrieving and adopting page rank and context tags algorithm using context information on the web 2.0 environment. A large-scale search engine's [15] in-depth

description and also described the page rank algorithm is detail. This algorithm stated that the page relevance is increase with the number of hyperlinks to it from the other page that relevant to that page. The details [23] of the Taxonomy and various process held of web mining. The beneficial point of view the defined [13] overview of web mining and their latest developments, trends with respect to web mining applications.

III SEARCHING AND RANKING FRAMEWORK

During the knowledge based process, the data pre-processing plays a crucial role and it can improve the quality of the data that helps to improve the accuracy and relevancy. In proposed work's architecture of search engine, pre-processing become important step during text based mining because the real-world data tend to be incomplete, inconsistent and dirty. In proposed search engine, all keywords with respect to their URLs has been stored in local database. The keyword as query entered in the search engine interface, after matching the keyword in local database all respective URLs has been being fetched. The new list of URLs has been construct against the keyword after the extracting keywords from each links as per the query entered. These all tasks of a search engine where request processed, extracted URLs against keyword and then stored in specific data structure are considered as the pre-processing. Afterword the stored list of URLs handed over to the ranking part of the search engine to rank the URLs as per the proposed algorithm approach.

During ranking, the proposed ranking algorithm is use to assign ranks to each of the URL. The process of rank a URL has been divided into major three 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 weight and all links of URL. The process of pre-processing and ranking the URLs as per the architecture shown in Fig - 1.

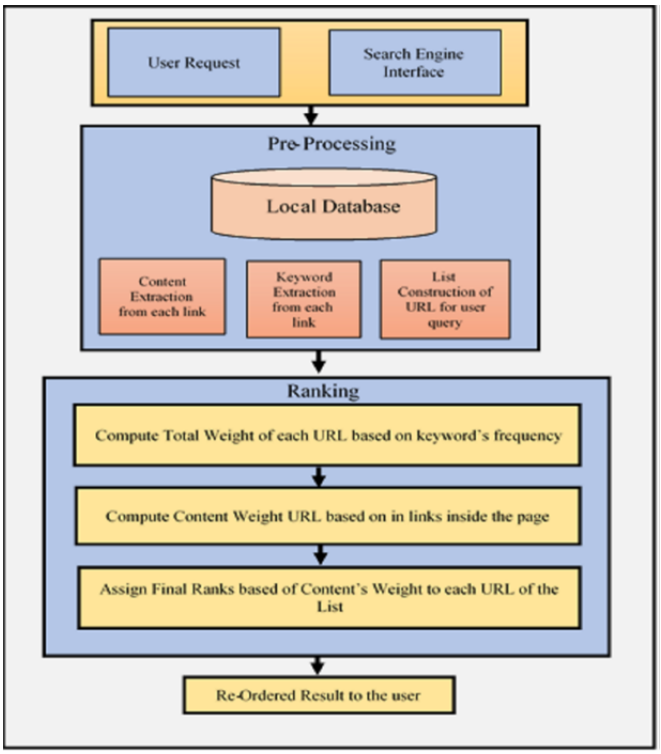

Figure 1: Architectural Design of Search Engine

After ranking algorithm, final rank has been assigned to each URL of the list. The list of URL and respective rank has been re-ordered and shown to the user as the result. The web page has highest rank is on the top and least at the bottom of the list.

IV ALGORITHM OF CONTENT'S WEIGHT BASED PAGE RANKING

Algorithm Name: Content's Weight based Page Ranking Algorithm
Input: A set URLs as *links*.
Output: Rank Score

(Pre-Processing)

1. User enters a keyword as Query for search (N)
2. Set an Array (X) of links {k=0 and nn= X. length} i.e. X= {$link_1$, $link_2$, $link_3$, ……, ……., $link_{nn}$} // Array of link
3. For each k<nn
   Set flag =0
4. If content is find in k_{th} index
   Set the flag=1 and status = True
5. While (status) index document// inner loop
6. Apply the following tasks on selected url

a) Remove stop words
b) Apply stemming on each line of content in page
c) Concatenate to String variable (Q_n)

7. If (Status==FALSE)
   Return false
   Else
   Z= strlen (Q_n) // Z= length of string
   LOOP START from 0 till of Z
   Search N & their synonyms in Q_n

(Ranking Part)

8. Find frequency of keyword & their synonyms (X_r)
   Set total_count = 0 and X_r= 0
   If ((strops (Q_n, N)! = false))
   {
   total_count=total_count+1
   }
   X_r=X_r+total_count
9. D=strlen($_SERVER['REQUEST_URI']))
   Here, D= No. of all query parameters passing from k_{th} url.
10. Total_Weight (T_w)= $\frac{X_r+D}{Z}$
11. Find all in-links (I_l) from k_{th} url:
    Set I_l=0
    If ((stropos (k_{th} ,"http")! =false) || (stropos (k_{th}, "https")! = false))
    {
    I_l=I_l+1
    }
12. To find all out-links (O_l) from k_{th} url:
    Call an inbuilt function "curl_init ($url)", with parameters i.e. k_{th} url & store returned value in O_l.
13. Find all External Image links (EI_l) from kth url:
    If ((strops (k_{th} ,"img")! =false)
    {
    EI_l=EI_l+1
    }
14. Calculate total no. of links (T_l) inside k_{th}url:
    $T_l = I_l + O_l + EI_l$
15. Find Content Weight (CW) of k_{th} url: for e.g. url *A*:
    CW(A)= I_l+ T_w
16. Final Content Weight based Page Rank (CWPR) of url *A*:
    CWPR(*A*)= $\sum \frac{CW(A)}{T_l}$
17. Set an Associative Array (AA) to store CWPR of *A* as well as URL.
18. If (k<nn)
    {
    goto Step 4
    Else
    Return false

}
LOOP END

19. Finally sort all URLs in an Array (AA) in descending order and show final result to user.

V EXPERIMENTAL RESULTS

For experimental results, a search engine has been developed followed by proposed algorithm. This search engine uses mathematical calculation to compute final page rank of web pages or URLs. After getting the content's frequency, computed its Content's weights and applying page rank using links, final rank of URL has been assigned to each URLs with respect to keyword as inputted query by the user. Experiment conducted with a keyword as a query of user 'mobile' against a particular search engine. Ten web pages taken as the database for search engine as an input shown in Table - I.

S.No	URL No	URL
1	UR 1	http://www.amazon.in/mobiles-phones/b?ie=UTF8&node= 1389401031
2	UR 2	http://in.priceprice.com/mobilephones/
3	UR 3	http://www.gsmarena.com/makers.php3
4	UR 4	https://www.infibeam.com/latest-mobile-phones
5	UR 5	https://www.ebay.in/cat/mobiles
6	UR 6	http://www.themobilestore.in/
7	UR 7	http://www.gadgetsnow.com/compare-mobile-phones
8	UR 8	http://gadgets.ndtv.com/mobiles/all-brands
9	UR 9	https://en.wikipedia.org/wiki/Mobile_phone
10	UR 10	https://www.snapdeal.com/products/mobiles-mobile-phones

TABLE 1: INPUT DATA SET

This input dataset has to analyzed on the basis of general search engine's searching technique approach. In this research work a search engine called 'Google' has been used to rank the web pages. The list of URLs has been re-ordered with respect to their ranks assigned by search engine as shown in Table -II.

TABLE II: RANKS BY GENERAL SEARCH ENGINE

S.No	URL No	URL	Search Engine (RANK)
1	UR 10	https://www.snapdeal.com/products/mobiles-mobile-phones	1
2	UR 8	http://gadgets.ndtv.com/mobiles/all-brands	2
3	UR 1	http://www.amazon.in/mobiles-phones/b?ie=UTF8&node=1389401031	3
4	UR 4	https://www.infibeam.com/latest-mobile-phones	4
5	UR 2	http://in.priceprice.com/mobilephones/	5
6	UR 5	https://www.ebay.in/cat/mobiles	6
7	UR 3	http://www.gsmarena.com/makers.php3	7
8	UR 7	http://www.gadgetsnow.com/compare-mobile-phones	8
9	UR 9	https://en.wikipedia.org/wiki/Mobile_phone	9
10	UR 6	http://www.themobilestore.in/	10

From the general search engine's results, it has been observed that ranks are assigned to each of the URLs. Highest rank web page among all the web page is shown on the top of the list i.e. 1st and least rank at the last position in the list i.e. 10th. For example, the UR 10 has the highest rank such as 1st according to an approach followed by particular search engine and UR 6 having least rank.

Finally, a search engine followed by proposed ranking algorithm called CWPRA has been applied on input dataset and got the updated list of URLs with new rank. The process of count keyword frequency, calculating composite content's weight of frequencies and assigned final page ranks with in-links has been done by using mathematically calculation at back-end. A numerical value get by mathematical calculation represents as their rank value. The sequence of URLs as input dataset has been change according to their ranks.

VI PERFORMANCE EVALUATION

In this work, sample dataset shown in table I considered for rank evaluation. Ranks computed from the general search engine shown in Table – II. Ranks will be computed and assigned manually using Content's Weight Based Page Rank algorithm to each URL as per mathematical calculation of algorithm. This dataset of 10 URLs given to the different users for their decision as per their relevancy and their decision has been shown in the form of ranks assigned to each URL as per their decisions. Now the same relevant dataset is evaluated against the dataset which is computed as per the CWPR algorithm and general search engine as shown in Table - III.

S.No	URL No	URL	General Search Engine	Manual Ranking	CWPRA
1	UR 10	https://www.snapdeal.com/products/mobiles-mobile-phones	1	6	6
2	UR 8	http://gadgets.ndtv.com/mobiles/all-brands	2	5	5
3	UR 1	http://www.amazon.in/mobiles-phones/b?ie=UTF8&node=1389401031	3	4	4
4	UR 4	https://www.infibeam.com/latest-mobile-phones	4	7	3
5	UR 2	http://in.priceprice.com/mobilephones/	5	10	10
6	UR 5	https://www.ebay.in/cat/mobiles	6	3	9
7	UR 3	http://www.gsmarena.com/makers.php3	7	8	8
8	UR 7	http://www.gadgetsnow.com/compare-mobile-phones	8	7	7
9	UR 9	https://en.wikipedia.org/wiki/Mobile_phone	9	2	2
10	UR 6	http://www.themobilestore.in/	10	1	1

TABLE III: COMPARISON OF RANKS

From the comparison table, it has been clearly shown that ranks with proposed approach become different from the other search engine's approach. On the same time the matching of manual ranking, general search engine ranking and CWPRA ranking has been done in table-III. As we can see that only two Documents UR- 4 and UR-5 are mismatched from manual ranking and CWPRA ranking whereas the ranks attained by general search engine does not match with the manual ranking that means the purposed ranking approach is more near to the user's interest and author can state that the proposed approach has the ability rank the web page on the basis of content's weight and links with more relevancy.

VII CONCLUSIONS AND FUTURE WORK

The search engine has been developed with established local database consists of keywords and their respective URLs. The proposed enhance approach for sure gives the different and relevant results as compare to general search engine's result with respect to combining parameters i.e. contents, its weight and page rank. Assigned ranks are numerical values based on mathematical calculation from the proposed ranking algorithm. Web page has their individual rank value on the basis of their importance which includes their content's frequency, weights and links of the web page.

As for future work, results will show on the basis of certain relevancy parameters such as Precision, Recall and Accuracy in upcoming research publication. Proposed methodology focus only on content and structure mining to rank the web page whereas relevant information or ranking important pages may be done with web usage mining also.

ACKNOWLEDGMENT

I would like to thanks Dr. Vijay Laxmi, Guru Kashi University for her spontaneous guidance and her valuable suggestion with respect to this paper. I also thanks to Dr. Arvinder Singh Kang for being open person to encouraging and shape my ideas for this research work. They help me in all stages of this research work and make happen this research work efficient.

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] C. Singh, V. Laxmi, and A.S. Kang, "A New Ranking Algorithm for Search Engine: Content's Weight based Page Ranking," International Journal of Computer Applications. 152(7) 0975 – 8887, 2016.

[5] 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.

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

[7] 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.

[8] 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.

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

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

[11] J. Gibson, B. Wellner, and S. Lubar, "Adaptive web-page content identification," Ninth annual ACM international workshop on Web information and data management (WIDM '07), New York, USA, 2007.

[12] B. Liu, and KC-C. Chang, "Editorial: Special issue on Web Content Mining," ACM SIGKDD Explorations Newsletter, New York, USA, 6(2) 1-4, 2004.

[13] G. Lappas, "An overview of web mining in societal benefit areas," Ninth IEEE International Conference on E-Commerce Technology, National Centre of Sciences, Tokyo, Japan, 2007.

[14] S. Chakrabarti, M. Berg, and B. Dom, "Focused Crawling: A New Approach to Topic-specific Web Resource Discovery," Journal Computer Networks: The International Journal of Computer and Telecommunications Networking, New York, NY, USA, 31(11-16) 1623-1640, 1999.

[15] S. Brin, and L. Page, "The anatomy of a large-scale hyper textual Web search engine," Computer Networks and ISDN Systems. Seventh International World Wide Web Conference, Brisbane, Australia, 1998.

[16] C. Wang, Y. Liu, L. Jian, and P. Zhang, "A Utility based Web Content Sensitivity Mining Approach," IEEE/WIC/ACM International Conference on Web Intelligent and Intelligent Agent Technology (WIIAT '08), University of Technology, Sydney, Australia, 2008.

[17] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: Information and pattern discovery on the World Wide Web," Ninth IEEE International Conference on Tools with Artificial Intelligence, (ICTAI '97), Newport Beach, California, 1997.

[18] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2001.

[19] M. Kobayashi, and K. Takeda, "Information Retrieval on the Web," Journal ACM Computing Surveys, New York, USA, 32(2) 144-173, 2000.

[20] J. Mayfield, and P. McName, "Indexing Using Both N-Grams and Words," In proceeding of NIST Special Publication 500 - 242: The Seventh Text Retrieval Conference (TREC '7), National Institute of Standards and Technology, Gaithersburg, MD, 1998.

[21] J. Cho, N. Shivakumar and H. Garcia-Molina, "Finding replicated web collections," ACM International conference on Management of data, (SIGMOD '2000), Dallas, TX, USA, 2000.

[22] S. Kim, and J. Kwon, "Information Retrieval using Context Information on the Web 2.0 Environment," International Journal of Computer Science and Network Security, 9(10), 2009.

[23] W. Bin, and L. Zhijing, "Web Mining research," 5th International Conference on computational Intelligence and Multimedia Applications, Washington, DC, USA, IEEE Computer Society, 2003.