View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 10 (2013) 773 – 780

Procedia Technology

1st International Conference on Computational Intelligence: Modeling, Techniques and Applications (CIMTA- 2013)

Analytical Design of Feature based Ranking

Sutirtha Kumar Guha[1,3,*] , Anirban Kundu[2,3], Somasree Bhadra[1,3], Rana Dattagupta[4]

[1]*Seacom Engineering College, Jaladhulagori, Sankrail, Howrah, West Bengal, India*
[2]*Kuang-Chi Institute of Advanced Technology, Shenzen, P.R China*
[3]*Innovation Research Lab, West Bengal, India*
[4]*Jadavpur University, West Bengal, India*

Abstract

In this paper, a new method is introduced to calculate the ranking of a Web-page in a Search Engine. We have considered inbound and outbound Web-page links, Session time spent in the Web-page, Relevancy of the Web-page with the searching string and Web Traffic for the Web-page as decisive factors. These factors are measured dynamically and generate a numeric value, termed as Decisive Factor (D_F). These Decisive Factors are calculated by newly introduced mathematical equations. Finally, all Decisive Factors are integrated to find the final ranking of the particular Web-pages. Since the access to a Web-page is often unpredictable, this type of Web-page ranking of dynamic nature is highly desirable to obtain more efficient and accurate Web-page ranking.

© 2013 The Authors. Published by Elsevier Ltd. Open access under CC BY-NC-ND license.

Selection and peer-review under responsibility of the University of Kalyani, Department of Computer Science & Engineering

Keywords: Web Traffic, Decisive Factor (D_F); Inlink; Outlink; Relevancy;

1. Introduction

Now-a-days all Search Engines use some kind of Web-page link-based ranking in their ranking algorithms. Without doubt, this has been the result of the success of Google, and its Web-page Ranking based link algorithm [1]. Taxonomy of different link ranking algorithms is presented in [2]. The typical idea behind Web-page ranking is that if Web-page 'A' has a link to Web-page 'B', then 'A' implicitly transfers some importance to 'B' [3]. Several techniques for ranking the Web-pages have been developed based on different criteria. No single method is adequate enough to perform the task efficiently. In a typical Search Engine, Web-pages are ranked based on specific features or by some predefined procedure according to the ranking of enlisted Web-pages. A Search Engine ranks the

*Corresponding author.
Sutirtha Kumar Guha. Tel.:+919804245316;
E-mail address: sutirthaguha@gmail.com.

2212-0173 © 2013 The Authors. Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
Selection and peer-review under responsibility of the University of Kalyani, Department of Computer Science & Engineering
doi:10.1016/j.protcy.2013.12.421

available / downloaded Web-pages in an order according to the relevance of the search query [4-5]. Web-page ranking is calculated based on the static parameters in case of typical Search Engines. The static methodology results in irrelevant Web-page ranking in most of the cases. So, top ranked Web-pages sometime may not consist of relevant information or may need more time to access due to overloaded Web-traffic. In both cases, the Web-page ranking may be considered less relevant. In our proposed work, the Search Engine based Web-page ranking is calculated in runtime. The ranking of a same Web-page can vary time to time based on the Web traffic on that particular time instance, the session time and other related parameters. Henceforth the calculated Web-page ranking is dynamic in nature. The unpredictable access to a Web-page causes this type of dynamic Web-page ranking more relevant and highly desirable with reference to the user.

This concept makes the Web-page ranking more relevant to the requirement compared to typical Web-page ranking methodology. Researchers from all over the world have worked on different sectors of Search Engine. Performance of Web-page ranking is increased by introducing hierarchical approach in [6]. The performance of the Web-page ranking can also be improved by implementing Cellular Automata concept in the working principle of Web-page ranking [7]. The data repository is supposed to work more efficiently and rapidly by implementing the distribution property [8]. A Virtual Feature based Logistic Regression (VFLR) ranking method is proposed to conduct the logistic regression on a set of essential independent variables, called virtual features (VF) [9]. A learning method is introduced to rank models named as RLM-AVF(Visual Feature based Ranking Learning Model) which is based on the large margin method, under the framework of structural SVM(Support Vector Machine) [10].
Rest of the paper is as follows: Proposed design is described in Section 2. Section 3 concludes the paper.

2. Proposed System Design

In proposed work, the Web-pages are ranked on basis of four attributes or features as follows:
(i) Inlinks and Outlinks of Web-page
(ii) Visitor's session
(iii) Relevancy of the Web-page based on searching string
(iv) Web traffic on a Web-page at any time instance

Different weightage is given to these above mentioned attributes. For each attribute, 'D_F' is calculated based on the proposed mathematical equations; and further final Web-page rank is calculated based on the different Decisive Factors.

2.1. Inlinks and Outlinks of Web-page

Inlinks are the Web-links of those Web-pages from which the current Web-page is re-directed; and, Outlinks are the Web-links of those Web-pages which are re-directed from the current Web-page. In our work, Web-page rank of all inlinked and outlinked Web-pages is taken into consideration as shown in Figure 1.

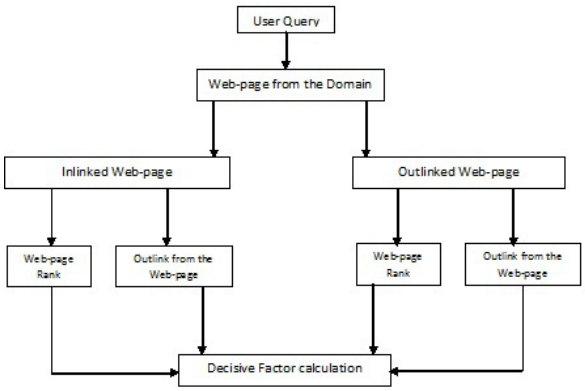

Figure 1. Operational Flowchart of Inlink and Outlink Decisive Factor Calculation

The 'Decisive Factor' is calculated based on Equation 1.

$$Inlink - Outlink\ Decisive\ Factor\left(D_{Fio}(A)\right) = \left(\sum (PR(Ti)) \div O(Ti)\right) \times \left(\sum (O(To) \div PR(To))\right) \quad (1)$$

where,
$D_F$io(A) = Inlink-outlink Decisive Factor of the Web-page 'A';
A = Current Web-page whose Decisive Factor would be calculated;
Ti = Web-pages which have a Web-link to Web-page A; i.e., inbound Web-pages or inlinked Web-pages;
To = Web-pages which are redirected from Web-page A; i.e., outbound Web-pages or outlinked Web-pages;
PR(Ti) = Web-page rank of the inbound Web-page Ti in Web-page A; O(Ti) = Number of outbound links from Web-page Ti;
PR(To) = Web-page rank of the outbound Web-page To from Web-page A; O(To) = Number of outbound links from Web-page To;
It is assumed that four Web-pages (A, B, C and D) are connected with each other by Inlinks and Outlinks as shown in Figure 2. Web-page 'A' has inbound link from Web-page 'C' and outbound links to Web-page 'B' and Web-page 'D'. The inter-connect ions of the Web-pages are represented in matrix format as shown in Table 1.

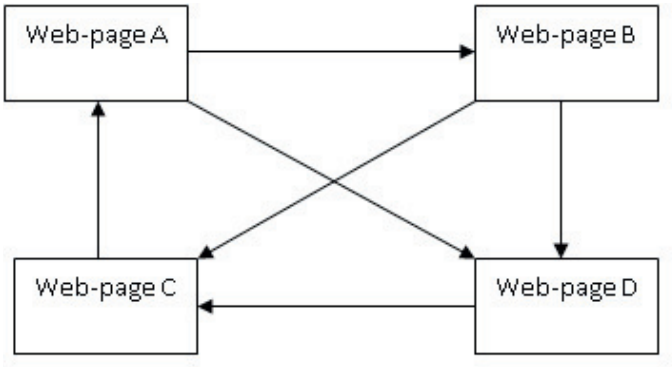

Figure 2. Control Flow Diagram of inter-connected Web-pages

Table 1. Inter-connectivity between Web-pages

Web-pages	A	B	C	D
A	X	√	X	√
B	X	X	√	√
C	√	X	X	X
D	X	X	√	X

According to our mathematical expression of Equation 1, calculated 'Decisive Factors' for the Web-page s A, B, C, D are as follows:
$D_F$io(A) = 3; $D_F$io(B) = 3; $D_F$io(C) = 1.66; $D_F$io(D) = 1.8;
In the above stated analysis, it is assumed that rank of a newly introduced Web-page would be '1'. Web - page rank is calculated based on the inlink-outlink Decisive Factor only. Hence, PR (Web-page) is replaced here by $D_F$io (Web-page).
Web-page 'C' is top ranked Web-page among the four (4) Web-pages according to the derived inlink-outlink 'Decisive Factor'.
Large number of inlinks of a particular Web-page indicates the relevancy and importance of that Web-page in the field. Similarly, outbound links indicate the less importance of the Web-page in the concerned field.

2.2. Visitor's Session

The duration of a visitor's session depends on the relevancy of the Web-page with respect to the requirement of the user. The total amount of 'time spent' of each visitor is calculated related to the specific Web-page taken into consideration. The Decisive Factor is calculated using the Equation 2 as follows:

$$Time\ Decisive\ Factor\left(D_{Ft}(A)\right) = \sum Vt \div \sum Vn \tag{2}$$

where,
Vt = Total time spent in the Web-page 'A' by a visitor;
Vn = Number of visitors of the Web-page 'A';

Vt and Vn are measured for any particular time; and, hence the value of Vt and Vn are always dynamic in nature based on user requirement at the particular time span. It is assumed in our paper that a Web-page may contain relevant information but the calculated time Decisive Factor may be poor due to the user interest at that moment.

This parameter reduces the ranking of the Web-pages having irrelevant contents. If a Web-page has unrelated data or contents, the user would not stay for a long time on that particular Web-page. Thus, session time would be less in case of unacceptable or poor result as compared to a useful Web-page. The operational behavior for calculating the 'Time Decisive Factor' is represented in Figure 3.

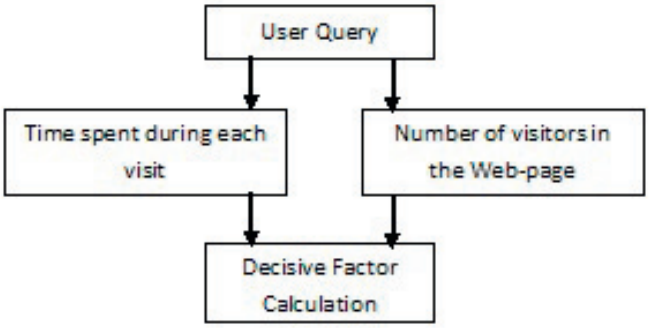

Figure 3. Operational Flowchart of Time Decisive Factor Calculation

Consider three cases as follows:

Case1:
Web-page 'A' is visited by 4 visitors and their spending times for the particular Web-page are 10 sec., 1 sec., 5 sec., 3 sec. respectively.
$D_Ft(A)) = \sum Vt/\sum Vn = (10+1+5+3)/(4) = 19/4 = 4.75$

Case2:
Web-page 'A' is visited by 7 visitors and their spending times for the Web-page are 1 sec., 2 sec., 3 sec., 4 sec., 1 sec., 2 sec., 1 sec. respectively.
$D_Ft(A)) = \sum Vt/\sum Vn = (1+2+3+4+1+2+1)/(7) = 14/7 = 2$

Case3:
Web-page 'A' is visited by 2 visitors and their spending times for the Web-page are 15 sec., 20 sec. respectively.
$D_Ft(A)) = \sum Vt/\sum Vn = (15+20)/(2) = 35/2 = 17.5$

According to the calculated 'Time Decisive Factor' value, Web-page 'A' is ranked high in 'Case 3'. It is assumed that high average session in a web page increases the Ranking of the particular Web-page.

2.3. Relevancy of Web-pages

Relevancy of the Web-pages with respect to the searching string is checked by the typical Semantic Web concept [3, 10]. The 'Relevancy Decisive Factor' is calculated using Equation 3.

$$D_{FI}(A) = \int_1^d Mf_i \tag{3}$$

where,

Mf = Successful Forward Movement of the Searching Pointer which is considered as the control that moves from one segment to another segment for continuation of the searching procedure.

Web-page relevancy is measured based on Ontological Semantic Web. A tree structured Search Engine database serves the purpose of implementing the Semantic Web concept in a typical manner. The operational control flow is shown in Figure 4.

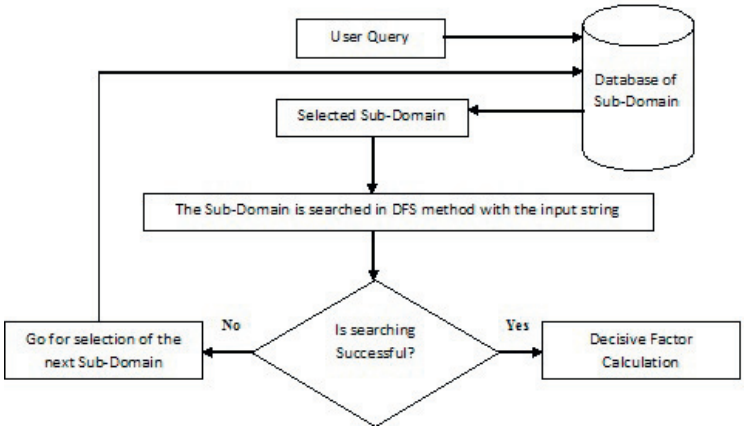

Figure 4. Operational Flowchart of Relevancy Decisive Factor Calculation

In case of finding relevancy of a Web-page, it is assumed that a typical Search Engine database contains the Web-page information in a hierarchical fashion. The searching for a specific domain is carried out in a Depth First Search (DFS) manner.

2.3.1. Analytical Study

In Ontological Semantic Web, the relevancy is measured using the specified relational path. The typical database of a Search Engine is shown in Figure 5. It is assumed that the database is categorized in domains like 'Environment', 'Sports', 'Education' and others. Each domain is again divided into several sub-domains. In Figure 5, sub-division of 'Education' domain is illustrated as taken input string is "Web- page ranking" related with educational field. The first domain 'Environment' is selected first and each sub - domain of 'Environment' is visited in DFS manner. I t is assumed that no matching field has been found in this 'Environment' domain since the input string is related with education field. If searching pointer does not find any appropriate domain after visiting to the end of the path, then it moves backward to the respective parent node which is the previous level node of the current node. The second domain 'Education' would be visited after the unsuccessful visit through the 'Environment' domain. Then, the input or searching string finds a suitable match in the 'Education' domain. The successful path would be "Education-

Science-Computer Science-Web Technology- Search Engine-Web-page Ranking" as shown in Figure 5. The 'Relevancy Decisive Factor' is calculated based on Equation 3.

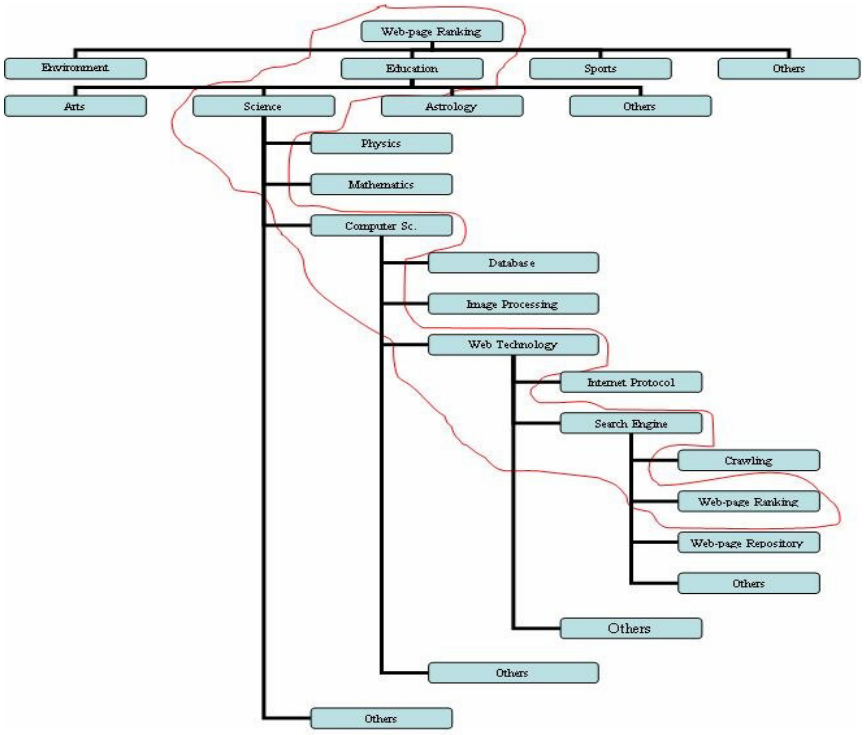

Figure 5. Diagram of a Hierarchical Database of a Search Engine

2.4. Web Traffic

In this paper, Web Traffic data is measured dynamically and the request from client is sent to the Server for processing information. It processes the data which is returned to the client as the list of matched URLs. It is presumed that client might request any URL at time t1 and the response is successfully achieved from that particular URL at time t2. It is expected that response delay might be occurred due to Web Traffic. Hence, response delay (t2 - t1) is considered as the parameter for Web Traffic. It is considered that there is a 'Threshold Value' (TV) for each Web-page based on the 'Web Traffic' concept. TV has a high range limit (TV (higher)) and a low range limit(TV (lower)). It is expected that the "Web Traffic" of any Web-page should be steady or slightly differ from its past record. The "Web Traffic" beyond the threshold rang e is considered as intrusion of some kind of unwanted noise within the system and it might increase or decrease the respective Web-page ranking. The TV is checked at each time instance for avoiding unwanted noise. Operational flowchart of "Web Traffic" is shown in Figure 6.

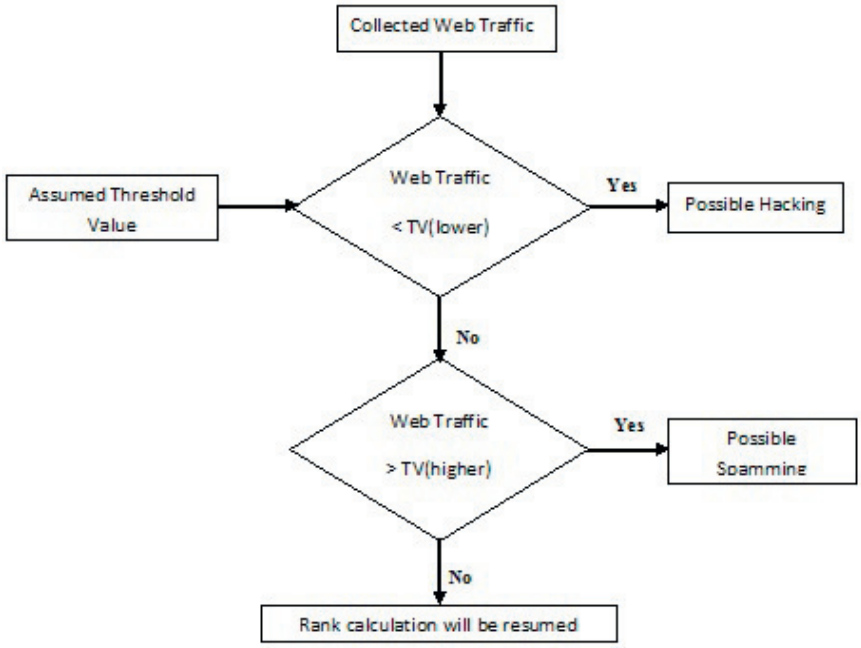

Figure 6. Operational Flowchart of Web Traffic Checking

2.5. Final Web-page Ranking Calculation

In the proposed work the resultant ranking of a Web-page is calculated based on five different D_F s as discussed in the previous sub-sections. Initially, a particular domain is selected using 'User Query'. Then, different types of Decisive Factors are being calculated using "Inlinks and Outlinks of Web-pages", "Visitor's session" and "Relevancy of Web-page". Finally, "Web Traffic" is utilized for ensuring the absence of any unwanted noise. In final calculation, different weightage is multiplied with the calculated "Decisive Factors" based on t heir importance to affect the overall Web-page ranking. The final Web-page ranking (PR) of a Web-page 'A' is as shown in Equation 4.

$$PR(A) = (((D_{Fi}(A) \times 0.5) + (D_{Ft}(A) \times 1) + (D_{FI}(A) \times 1.5)) \div 3) \quad (4)$$

2.5.1. Analytical Study

Case1:
Let, in any time instance the calculated Decisive Factor values for a Web-page 'X' are as follows:
$D_F$io(X) = 2.5
D_Ft(X) = 2.5
D_Fr(X) = 2.5
Then, PR(X) = (((2.5 * 0.5) + (2.5 * 1) + (2.5 * 1.5)) / 3) = 1.25 + 2.5 + 3.75 = 7.5

Case2:
Let, in any time instance the calculated Decisive Factor values for a Web-page 'X' are as follows:
$D_F$io(X) = 3.5
D_Ft(X) = 2.5
D_F r(X) = 2.5
Then, PR(X) = (((3.5 * 0.5) + (2.5 * 1) + (2.5 * 1.5)) / 3) = 1.75 + 2.5 + 3.75 = 8.0

Case3:
Let, any time instance the calculated Decisive Factor values for a Web-page 'X' are as follows:
D_F io(X) = 2.5
D_Ft(X) = 3.5
D_Fr(X) = 2.5
Then, PR(X) = (((2.5 * 0.5) + (3.5 * 1) + (2.5 * 1.5)) / 3) = 1.25 + 3.5 + 3.75 = 8.5

Case4:
Let, any time instance the calculated Decisive Factor values for a Web-page 'X' are as follows: $D_F$io(X) = 2.5
D_Ft(X) = 2.5 D_Fr(X) = 3.5
Then, PR(X) = (((2.5 * 0.5) + (2.5 * 1) + (3.5 * 1.5)) / 3) = 1.25 + 2.5 + 5.25 = 9.0

It is derived from different case studies as discussed above that increasing the 'Decisive Factor' values by same amount does not increase the Web-page rank uniformly. It depends on the multiplied weightage of that specified 'Decisive Factor' value. The value of multiplied weightage varies based on the importance of the particular Decisive Factor. In our proposed work it is assumed that relevancy of a web page with the searching string should be the most important Decisive Factor for calculating the web page ranking, hence that Decisive Factor is multiplied by highest valued weightage. The weightage is introduced to assign different importance to the Decisive Factors. Hence, the Web-page ranking would be higher in case of changing the Decisive Factor with higher multiplied weightage and vice-versa.

3. Conclusion

In this paper, a new approach for calculating the Web-page ranking of a typical Search Engine is introduced. Different features of a Web-page are considered as "Decisive Factors" (D_Fs) which are responsible for changing the rank of a Web-page in the Search Engine. Different "Decisive Factors" are calculated for the Web-pages based on the proposed mathematical equations. Four D_Fs are used for calculation of the 'Decisive Factors'. The calculated 'Decisive Factors' have been assigned with different weightage. The 'Decisive Factors' and assigned weightage finally calculate the rank of the Web-page.

References

[1] S. Brin and L. Page; "The anatomy of a large-scale hyper textual Web Search Engine"; In 7th WWW Conference, Brisbane, Australia, April 1998.
[2] C. Ding, X. He, P. Husbands, H. Zha, H. Simon; " page rank, hits and a unified framework for link analysis"; LBNL Tech Report 49372, 2001-2002.
[3] Shuming Shi, Jin Yu, Guangwen Yang, Dingxing Wang; "Distributed Page Ranking in Structured P2P Networks"; ICPP 2003.
[4] Debajyoti Mukhopadhyay, Debasis Giri, Sanasam Ranbir Singh; "An approach to confidence based page ranking for user oriented Web search"; SIGMOD Record 32(2): 28-33, 2003.
[5] Debajyoti Mukhopadhyay, Pradipta Biswas; "FlexiR ank: An Algorithm Offering Flexibility and Accuracy for Ranking the Web Pages"; International Conference on Distributed Computing & Internet Technology, ICDCIT 2005 Proceedings, Bhubaneswar, India; Lecture Notes in Computer Science Series, Springer-Verlag, Germany; December 22-24, 2005.
[6] Debajyoti Mukhopadhyay, Pradipta Biswas, Young-Chon Kim; "A Syntactic Classification based Web Page Ranking Algorithm"; The 6th International Workshop MSPT 2006 Proceedings; Youngil Publication; ISBN 89-8801-90-0, ISSN 1975-5635; Republic of Korea; November 20, 2006.
[7] A. Kundu, R. Dutta, D. Mukhopadhyay; "Converging Cellular Automata Techniques with Web Search Methods to Offer A New Way to Rank Hyper-Linked Web-Pages"; International Symposium on Information Technology Convergence (ISITC 2007); Jeonju, Republic of Korea; IEEE CPS Publishing Services; November 23-24, 2007.
[8] A. Kundu, R. Dutta, D. Mukhopadhyay; "An Alterna te Way to Rank Hyper-linked Web Pages"; 9th International Conference on Information Technology (ICIT 2006); Bhubaneswar, India; IEEE Computer Society Press, New York, USA; December 18-21, 2006.
[9] Cai F, Guo D, Chen H, Shu Z, "Your Relevance Feedback Is Essential: Enhancing the Learning to Rank Using the Virtual Feature Based
[10] Logistic Regression"; PLos ONE7(12): e50112. doi: 10.1371/journal.pone.0050112, 2012
[11] Xia Li, Jianjun Yu, Jing Li, "RLM-AVF: Towards Visual Features Based Ranking Learning Model for Image Search", Electrical Engineering and Control (Springer), Volume 98, pp 135-140, 2011.