A Survey of various Web Page Ranking Algorithms

Mayuri Shinde
Research Scholar,
Department of Information Technology
Maharashtra Institute of Technology Pune 411038, India

Sheetal Girase
Assistance Professor,
Department of Information Technology
Maharashtra Institute of Technology Pune 411038, India

ABSTRACT

Identification of opinion leader is very important in this world of internet because with the identified opinion leaders in any application area such as Knowledge related sites, followers or other individuals can get valuable information more efficiently through direct communication with opinion leader. Internet i.e. WWW (World Wide Web) is the huge and very popular way of information broadcasting and communication. This huge www has so many web structures and within one structure there would be millions of web resources (contents, links) may exist. There are large numbers of webpages on the web which are linked to each other through hyperlinks. So, graph based techniques can be used to identify opinion leader i.e. techniques for ranking the results to provide the "best" results first. Different algorithms are there which are used for link analysis i.e. for ranking the web pages like PageRank (PR), Weighted PageRank (WPR), Hyperlink-Induced Topic Search (HITS), Spamming Resistant Expertise Analysis and Ranking (SPEAR) etc. This paper is focused on the study of different ranking techniques. Further this paper shows advantages, limitations and comparison of these techniques.

General Terms

Data and Web Mining.

Keywords

WWW (World Wide Web), Opinion leader, PageRank (PR), Weighted PageRank (WPR), Hyperlink-Induced Topic Search (HITS), Spamming Resistant Expertise Analysis and Ranking (SPEAR) .

1. INTRODUCTION

Opinion leader is the one who is expert in particular domain, he has that ability to influence others opinions, guide them.[13][14][15] So identification of opinion leader is very important. There are many techniques for identification of opinion leader.[16][17] One of them is identification using graph based approach. So this paper is survey of graph based techniques i.e. algorithms used for ranking web pages.

WWW contains very large hyperlinked and non-homogeneous information including text, image, audio, video, and metadata. It is becoming difficult to manage the information on the web and satisfy the user needs, with the rapid growth of information sources available on the WWW and growing needs of users. In order to find, extract, filter and order the desired information, it has become increasingly necessary for users to use some information retrieval techniques. Generally most of the users use information retrieval tools like search engines to find information from the WWW. Some ranking mechanism (web mining) either in back ends or in front end is used by most of the search engines before representing the pages to the user.

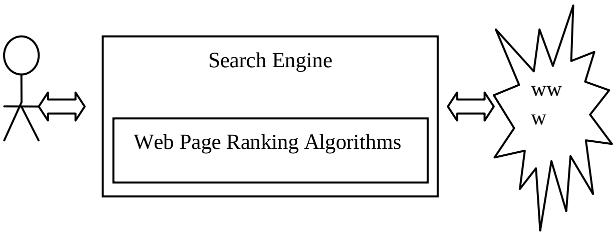

Fig 1: Concept of search engine

This picture (figure 1) show only about the basic idea behind search engine, there are many other things also there in this searching technique like, crawler, indexer, query processor etc. But this paper is a survey of page ranking algorithms. So It do not discuss these things but in this survey, it will cover page ranking algorithms and its variations.

2. WEB MINING

Data Mining is the process of Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from large databases. Web Mining is the application of data mining technique which is used to discover and retrieve useful information (knowledge) from the WWW document and services. Web mining can be divided into three categories namely web content mining, web structure mining and web usage mining as shown in Fig 2.[1][7]

When Huge amount of web data, information, files on the internet throughout the world is there, web mining came into picture. Mining of data present in the World Wide Web database in the form of web pages is called as Web mining. Due to this, user gets relevant information from the web. It is used to discover the content of the Web, the users' behavior in the past, and the webpages that the users want to view in the future. Hence in web mining different techniques of data mining are applied on the web data's. Web data can be:

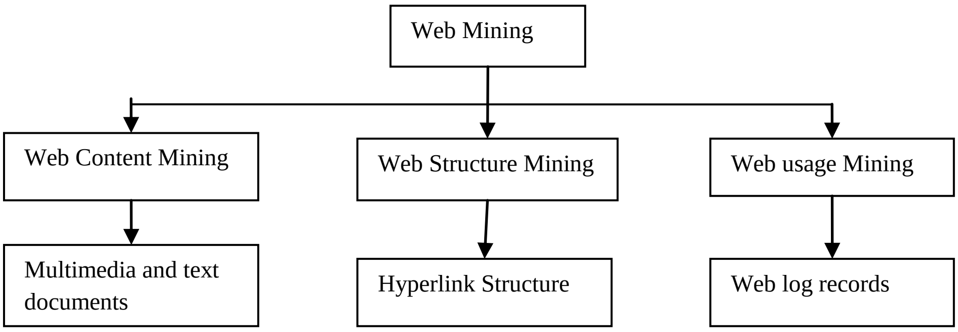

Fig 2: Types of Web Mining

- web pages contents like text and images
- HTML tags or XML tags i.e. Intra page structure.
- link structure between Web pages i.e. Inter page structure.
- data, which describe how web pages are accessed by various visitors on the internet i.e. usage data.

2.1 Web Content Mining

Web Content Mining (WCM) means examine the content of web pages and also result of web Searching. The content include text as well as graphics data. Web content mining has two types-

Web page content mining and search results mining

- It can be applied on web pages itself or on the result pages obtained from a search engine.
- traditional searching of web pages with the help of web pages is web page content mining.
- For efficiency, effectiveness and scalability, Web content mining uses data mining techniques.

2.2 Web Structure Mining

Web Structure Mining (WSM) tries to discover the link structure of the hyperlinks at the inter-document level unlike WCM which focuses on the structure of inner-document. It generates graph i.e. used to generate structural summary about the web pages where web pages act as nodes and hyperlinks as edges connecting two related pages.

- using this, information is obtained from actual structure of pages on the web.
- uses of Web Structure Mining:

  -Creating similarity measure between documentsClassifying Web pages
- Popular techniques used for Web Structure Mining:

  -Page RankHITS

2.3 Web Usage Mining

- To discover user navigation patterns and the useful information from the web data present in server logs, which are maintained during the interaction of the users while surfing on the web, WUM is used **.**
- It can be further categorized in finding the general access patterns or in finding the patterns matching the specified parameters.
- It is technique to predict user behaviour when it is interact with the web.

3. WEB PAGE RANKING ALGORITHMS

The web content mining (WCM) mainly concentrates on the document structure whereas web structure mining (WSM) explore the link structure inside the hyperlink between different documents and classify the web pages.

WSM is seen as an important approach to web mining , as the numbers of inlinks i.e. links to a page and of outlinks i.e. links from a page are valuable information in area of web mining.

There are various challenges associated with the ranking of web pages such that some web pages are designed only for navigation purpose and some pages of the web do not possess the quality of self descriptiveness. For ranking web pages, several algorithms were proposed in the literature.

In the following, we will present and discuss two important algorithms used for ranking web pages and their variations.

- PageRank
- HITS

3.1 PageRank

History of PageRank algorithm

PageRank is developed in 1996 at Standford university by Larry Page and Sergey Brin working on the Stanford Digital Library Project (SDLP). “To develop the enabling technologies for a single, integrated and universal digital library" was the goal of SDLP and was funded through the National Science Foundation among other federal agencies. [2]

In search for a dissertation theme, Page considered among other things exploring the mathematical properties of the World Wide Web, understanding its link structure as a huge graph. Terry Winograd who was his supervisor encouraged him to pick this idea (which Page later recalled as "the best advice I ever got") and Page focused on the problem of finding out which web pages link to a given page, considering the number and nature of such backlinks to be valuable information about that page [http://en.wikipedia.org/wiki/PageRank].

PageRank

The most commonly used algorithm for ranking various pages is Page Rank algorithm. GOOGLE which is very popular search engine of today uses PageRank that depends upon the link structure of the web to determine the importance of web pages. To order its search results GOOGLE uses PageRank so that documents that are seem more important move up in the results of a search accordingly. The working behind the PageRank algorithm is that a page with a large number of in-links(a link from an important page) to it, then its outgoing links to other pages also become important. It gives more importance to back links of a web page and propagates the ranking through links.

Thus, the page is important if it obtains a high rank i.e. if the sum of the ranks of its backlinks is high. Although many factors are considered while determining the overall rank but PageRank algorithm is the heart of Google

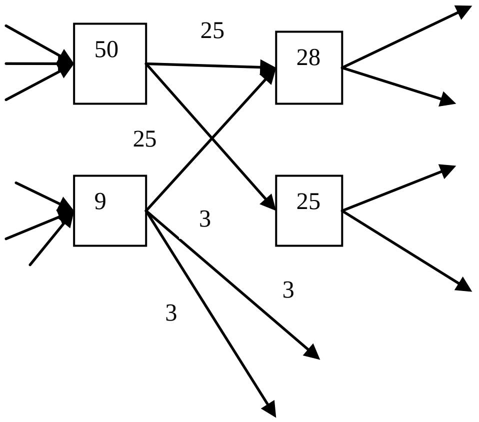

Fig 3: Simplified PageRank calculation

The links from one page to another is considered as a vote. The number of votes that a page receives is important and the importance of pages that casts the vote is also important. PageRank take pages as nodes and hyperlinks as edges of a graph. Each node i.e. web page has its own Rank value and distributes it evenly to its neighbors as shown in above fig.3

Existence of a link from one page to another page may indicate that the author of is interested in page. Following equation shows simplified version of PageRank [3]:

$$PR(A)=(1-d)+d(PR(T1)/C(T1)+\ldots..+PR(Tn/C(Tn))$$

$$\sum PR(T)=1$$

Where,

A is a web page,

PR(A) is rank score of page A,

PR(Ti) is the PageRank of the Pages Ti which links to page A,

C(Ti) denotes the number of outgoing links on page Ti ,

d is damping factor

The PageRank forms a probability distribution over the web pages. So, the sum of PageRanks of all web pages will be one. PageRank of a page depends on the number of pages pointing to a page i.e. inlinks.

The problem with above equation is that PR(Ti) values, the PageRanks of pages inlinking to page A, are unknown. To deal with this problem, Brin and Page used an iterative procedure. That is, they assumed that, in the beginning, all pages have equal PageRank. Now the rule in above equation is followed to computer PR(A) for each page in the index. The rule in above equation is successively applied, substituting the value of the previous iterate into PR(T). Iterative procedure is repeated with the hope that the PageRank scores will eventually converge to some final stable values. Above Equation compute PageRank one page at a time.

d is the damping factor. This damping factor d is important because users will only continue clicking on links for a finite amount of time before they get distracted and sometimes start exploring something completely unrelated. With the remaining probability (1- d), the user will click on one of the links on page A at random. Damping factor is usually set to 0.85. So it is easy to infer that every page distributes 85% of its original PageRank evenly among all pages to which it points i.e. total vote is “damped down” by multiplying it to 0.85.

The problem with this algorithm is that it is highly susceptible to spamming and does not favor important pages with only a few in-links even if that page is more important.

Problems of PageRank Algorithm are:

- It is a static algorithm, because popular pages tend to stay popular generally.
- Sometimes, Popularity of a site does not guarantee the desired, specific information to the searcher so another relevance factor also needs to be included.
- Algorithm is not fast enough and in Internet, available data is huge.

3.1.1 Weighted PageRank Algorithm

If the page is more popular then it means that it is having more webpages which are pointing i.e. linking to it. Wenpu Xing and Ali Ghorbani proposed a extended PageRank algorithm–Weighted PageRank algorithm-which assigns larger rank values to more popular pages instead of dividing the rank value of a page evenly among its outlink pages. Each outgoing link(page) get a value proportional to its importance i.e. its number of inlinks and outlinks. In this algorithm, weight is assigned to both inlinks and out links. Inlink is defined as number of links points to that particular page and out link is defined as number of links goes out from that particular page. This algorithm is more efficient than PageRank algorithm because it uses two parameters i.e. backlink (inlinks) and forward link (outlinks). The popularity (importance) of a page from the number of inlinks and outlinks is recorded a $\boldsymbol{W_{(v,u)}^{in}}$ and $\boldsymbol{W_{(v,u)}^{out}}$, respectively. $\boldsymbol{W_{(v,u)}^{in}}$ is the weight of *link*(*v*, *u*) which is calculated based on the number of inlinks of page *u* and the number of inlinks of all reference pages of page *v*. [4]

$$\mathrm{W}_{(\mathrm{v,u})}^{\mathrm{in}} = \frac{\mathrm{Iu}}{\sum_{\mathrm{p€R(v)}} \mathrm{Ip}}$$

Where

Iu is the number of inlinks of page u

Ip represent the number of inlinks of page p,

R(v) denotes the reference page list of page v

$W_{(v,u)}^{out}$ is the weight of link(v, u) which is calculated based on the number of outlinks of page u and the number of outlinks of all reference pages of page v.[4]

$$\mathrm{W}_{(\mathrm{v,u})}^{\mathrm{out}} = \frac{\mathrm{Ou}}{\sum_{\mathrm{p€R(v)}} \mathrm{Op}}$$

where

Ou is the number of outlinks of page *u*

Op is the number of outlinks of page *p*,

R(*v*) denotes the reference page list of page *v*.

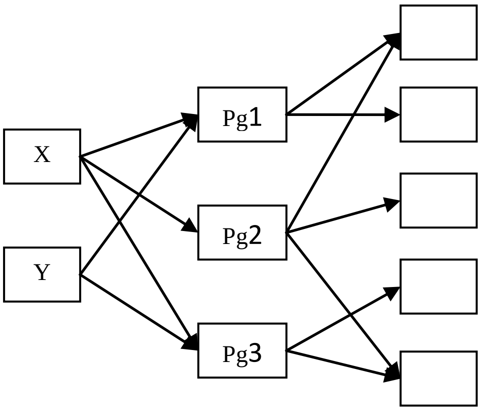

Fig 4: Example of Weighted PageRank Algorithm

Consider above diagram, In this example, Page A has three reference pages: pg1,pg2 and pg3. The inlinks and outlinks of these two pages are Ipg1 = 2, Ipg2 = 1, Opg1 = 2, and Opg2 = 3.Therefore,

$$W^{in}_{(X,Pg1)} = \frac{I_{pg1}}{I_{pg1} + I_{pg2}} = 2/3$$

And

$$W^{out}_{(X,pg1)} = \frac{O_{pg1}}{O_{pg1} + O_{pg2}} = 2/5$$

Considering the importance of pages, the original PageRank formula is modified as [4]

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

For the comparison of the WPR with the PageRank, the resultant pages of a query are categorized into four categories based on their relevancy to the given query. They are:

1. Very Relevant Pages (VR): Very important information related to a given query is contained in these pages.
2. Relevant Pages (R): These Pages are relevant but not having important information about a given query.
3. Weakly Relevant Pages (WR): These Pages may have the query keywords but they do not have the relevant information.
4. Irrelevant Pages (IR): These Pages are not having any relevant information and query keywords.

Relevancy Rule: The Relevancy of a page to a given query depends on its category and its position in the page-list. The larger the relevancy value, the better is the result.

$$K = \sum_{i€R(p)} (\mathrm{n} - \mathrm{i}) \times \mathrm{Wi}$$

Where

i represents the ith page in the result page-list R(p),

n represents the first n pages chosen from the list R(p),

and Wi is the weight of ith page.

Wi = (v1,v2,v3,v4)

Where

v1, v2, v3, v4 - value assigned to a page if the page is **VR, R, WR** and **IR** respectively.

The values are always v1>v2>v3>v4.

3.1.2 TrustRank Algorithm

TrustRank is a link analysis technique which is used for semi-automatically separating useful webpages from spam. For misleading search engines i.e. for obtaining high rank, many web pages are created. These pages are chiefly created for commercial reasons. They use different techniques to achieve high rank. It is true that human experts can easily identify spam, but it is too expensive to manually evaluate a large number of pages. TrustRank algorithm relies on PageRank [5].

TrustRank method first selects a small set of seed pages which is to be evaluated by an expert. . Assumption behind this is, if page is further away from good seed pages, then it has less probability that the page is good. The researchers who proposed the TrustRank methodology have continued to refine their work by evaluating related topics, such as measuring spam mass. [6]

TrustRank is used to separate reputable websites from spam and it is semi-automatic algorithm. It mainly based on the concept of trust attenuation.

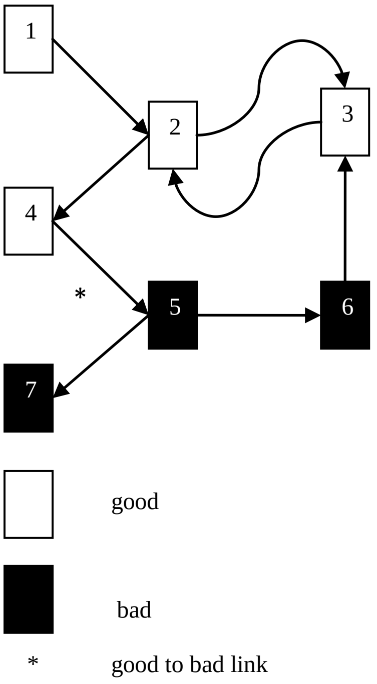

Fig. 5: A web of good (white) and bad (black) nodes

Consider above diagram i.e. Fig. 5 [3], Consider above diagram, the pages that are at most 2 links away from the good seed pages are page 2 and page 4. The probability that we reach a good page in at most 2 steps is 1, as both of them are good. The number of pages reachable from the good seed in at most 3 steps is 3. But only two of them (pages 2 and 4) are good and page 5 is bad. So, the probability of finding a good page is 2/3. And due to this, trust is reduced as one move further away from good seed pages.

It is shown that this algorithm can effectively identify a significant number of non-spam pages. [3][5] It can be used in search engines either in combination with Page Rank to rank results or separately to filter the index.

TrustRank assign a core vote of trust to a seed set to help search engines identify useful and spam pages. Through link from the seed sites this trust is attenuated to other sites.

3.2 HITS

WSM based algorithm called Hyperlink-Induced Topic Search (HITS) was developed by Kleinberg [8]. This algorithm is very popular and effective and used for rating and ranking websites, documents based on the link information. HITS algorithm process inlinks and outlinks of web pages for ranking. This algorithm allocates two attributes to a page, authority score and hub score. HITS assume that a good hub is a document that points to many other hyperlinks, and a good authority is a document which is pointed by many hyper links.

Following diagram illustrate HUB and authority[8]:

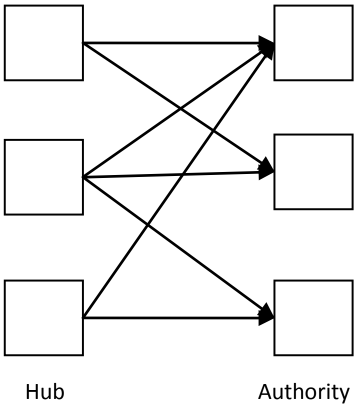

Fig. 6: Hubs and Authorities

HITS algorithm performs series of iterations; each iteration consists of following two basic steps:

- Authority Update Rule: Update the authority score or weight of each page to be proportional to the sum of hub weights of each page that link to it. So, high authority is given to a page if the pages which are taken as hubs for information link to them.

We update authority of page p to summation of hub:

$$auth(p) = \sum_{i=1}^{n} hub(i)$$

Where
n represents total number of pages that are
connected to p.

- Hub Update Rule: Update the Hub score of each page to be proportional to sum of authority weights of pages that it links to. So, high hub is given to a page if the pages which are taken as Authorities on subject link to them.

We update hub score of page p to summation of
authorities:

$$hub(p) = \sum_{i=1}^{n} auth(i)$$

Where
n represents total number of pages p connects to and I
is a page while p connects to.

HITS algorithm which is link based, rank web pages and this is done by analyzing their textual contents against a given query and after collecting the web pages i.e. a focused subgraph or base set, this algorithm consider only the structure of the web and not their textual contents.[9] HITS assign page as a hub that act as resource lists and it assign authorities which are has important contents. Sometimes a page may behave as hub as well as authority. It is executed at query time i.e. query time processing not at the time of indexing.

Some of limitations of original HITS algorithm

- Making difference between hubs and authorities is not easy because many sites act as both i.e. hubs as well as authorities.
- Because of equivalent weights, sometimes it happens that HITS may not generate the most relevant or wanted documents to the user queries.
- For the links which are automatically generated and which May not have relevant topics for the user query, HITS gives equal importance to them i.e. assume equal weight.
- If root set is not appropriate then in the result list, some queries can return non relevant documents and it can give false results.

To deal with some of these constraints, some improvements are done in original HITS algorithm.[10]

3.2.1 Spear (Spamming-Resistant Expertise Analysis And Ranking)

SPEAR is a graph- based ranking algorithm which is a new technique to define and find expertise of users which depends upon users activities. The main goal of SPEAR is to identify expert who has ability to find new and high quality information on Web.

The SPEAR algorithm relies on HITS, but it differs from the HITS algorithm because of two aspects [11]:

- Two sets i.e. a set of users and a set of documents are taken into consideration instead of a single set of documents, and the number of users and the number of documents under consideration does not necessarily equal. So, the adjacency matrix is not a square matrix.
- Depending on the user’s activity, for example- tagging), initialize the matrix with different values instead of 1 or 0 for the cells in adjacency matrix.

However, It is proved that SPEAR is guaranteed to converge in the same way as HITS.

For calculating the confidence i.e. level (knowledge) of a certain user in a specific skill SPEAR algorithms can be used.[12]. To introduce concept of expert, the SPEAR algorithm [11], use HITS. Expert is the one who has High level knowledge, skill in particular domain.

SPEAR is depends on two assumptions for this and these are
main elements of this algorithm. :

1. Mutual reinforcement of user expertise and document quality: Expertise of user in any particular topic is mainly depends upon documents quality which is found by user and similarly, the quality of document is depends on the user expertise who have found them i.e. expert should be selected depending on quality, not only quantity.
2. Discoverers vs. followers: Discoverer is the expert which finds -quality and relevant information. And follower is a user that annotates a document after a discoverer does.

4. CONCLUSION

This paper gives the basic idea and introduction of web mining and its related techniques i.e. types such as web content mining, web structure mining and web usage mining. Web structure mining algorithms or techniques can be used for opinion leader identification, as these techniques are used to rank results. User needs relevant information to cater to their needs. Therefore, it is very important to find the Web content and retrieve the user's interests and needs. Different algorithms are there which are used for link analysis like PageRank (PR), Hyperlink-Induced Topic Search (HITS) algorithms and their variations. In future, these algorithms can be used along with different techniques to identify and rank opinion leader and get relevant information.

5. REFERENCES

[1] Neelam Duhan, A. K. Sharma, Komal Kumar Bhatia," Page Ranking Algorithms: A Survey", 2009 IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009.

[2] Kaushal Kumar, Abhaya, Fungayi Donewell Mukoko, "PageRank algorithm and its variations: A Survey report", Sep. - Oct. 2013, OSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 1, PP 38-45

[3] MRIDULA BATRA, SACHIN SHARMA, "COMPARATIVE STUDY OF PAGE RANK ALGORITHM WITH DIFFERENT RANKING ALGORITHMS ADOPTED BY SEARCH ENGINE FOR WEBSITE RANKING", Sachin Sharma et al ,Int.J.Computer Technology & Applications,Vol 4 (1), 8-18, Jan-Feb 2013, ISSN:2229-6093

[4] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04), 2004 IEEE

[5] Zolt´an Gy¨ongyi, Hector Garcia-Molina, Jan Pedersen, "Combating Web Spam with TrustRank", Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004

[6] Vijay Krishnan, Rashmi Raj, "Web Spam Detection with AntiTrust Rank".

[7] T.Nithya, "Link Analysis Algorithm for Web Structure Mining", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 8, August 2013

[8] Kleinberg J., "Authorative Sources in a Hyperlinked Environment". Proceedings of the 23rd annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998.

[9] Bouchra Frikh , Brahim Ouhbi, Amine Ameur, "A Comparative Study of link analysis Algorithms for Information Retrieval", 2012 Next Generation Networks and Services NGNS, 2-4 December 2012 Portugal

[10] Xianchao Zhang, Hong Yu, Cong Zhang, and Xinyue Liu, "An Improved Weighted HITS Algorithm Based on Similarity and Popularity", Second International Multisymposium on Computer and Computational Sciences, 2007 IEEE

[11] Ching-man Au Yeung, Michael G. Noll, Nicholas Gibbins, Christoph Meinel, Nigel Shadbolt, " SPEAR: SPAMMING-RESISTANT EXPERTISE ANALYSIS AND RANKING IN COLLABORATIVE TAGGING SYSTEMS", Computational Intelligence, Volume 99, Number 000, 2009

[12] Jose María Álvarez-Rodríguez, Ricardo Colomo-Palacios, "Assesing professional skills in a multi-scale environment by means of graph-based algorithms", 2014 European Network Intelligence Conference, 2014 IEEE

[13] Lincheng Jiang, Bin Ge, Weidong Xiao, Mingze Gao, "BBS Opinion Leader Mining Based on An Improved PageRank Algorithm Using MapReduce," 2013 IEEE.

[14] Luo Jing, Xu Lizhen," Identification of Microblog Opinion Leader Based on User Feature and Interaction Network," 2014 11th Web Information System and Application Conference, 2014 IEEE.

[15] Do Kyun Kim, Anita C. James, Gregory J. Shepherd, A dissertation-"Identifying Opinion Leaders by Using Social Network Analysis: A Synthesis of Opinion Leadership Data Collection Methods and Instruments," August 2007.

[16] Haseena Rahmath P," Opinion Mining and Sentiment Analysis - Challenges and Applications," International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 3, Issue 5, May 2014, ISSN 2319 – 4847.

[17] Xiaofei Zhang & Dahai Dong," Ways of Identifying the Opinion Leaders in Virtual Communities," International Journal of Business and Management, Vol. 3, No. 7 July, 2008.