Community Influence Analysis in Social Network

Ximing Li
Villanova University

Table of Contents

Abstract

Online Social Networks (OSNs) have gained utmost popularity during current epoch. Users may get the information of their interests via variety of social media, such as Facebook, Twitter and DBLP. A group of users with common interests form a community that attracts other users towards itself to share their views and information, hence, community is an important structure within OSNs. Therefore, identifying the most influential community in OSNs arises.

In this project, a mathematical definition of community influence analysis and the corresponding modified PageRank algorithm is proposed. The modification includes the internal and external citation weight of different venues. Hence, this modification incorporates a better aspect to evaluate the real impact of each community. Furthermore, it identifies the most influential community in scholastic OSNs. It is the first time to analyse the community level influence with such a large scope. The proposed solution considers the impact of internal and external influential factors of communities upon each other. Subsequently, these factors are considered for introducing a weightage formula to calculate each community's joint weight. The citation network of distinct scholastic communities is considered for acquiring the citation information. Extensive comparative experiments are performed to evaluate the algorithm's performance with variant damping factors.

Chapter 1: Introduction

Internet is providing the opportunity to the users for utilizing it as the foremost interactive medium. During the last decades, we have witnessed a massive transition in the applications and services hosted on the Web. The social and participatory characteristics that are included in these services not only allow the users to obtain information but also provide the opportunities to actively participate for generating the contents. Theses incredible services lead toward the generation of virtual communities, where user could share their ideas, knowledge, experience, opinions and even media contents.

With the pervasive presence and ease of the user of the Web 2.0, an increasing number of people with different backgrounds, flock to the web to for enjoying many previously inconceivable activities[8] . The explosive of growth of social media has provided millions of people the opportunity to create and share content on a scale barely imaginable a few years ago. Massive participation in these social networks is reflected in the countless number of opinions, news and product reviews that are constantly posted and discussed in social sites.

Online Social Networks (OSNs) have gain immense esteem as well as experienced fast growth during the past decade. Leading actors in OSNs including Sina Weibo, Twitter, Google Scholar, DBLP, ArnetMiner etc. share a bulk quantity of users. Nowadays, massive users share much information about themselves and conduct a lot of activities utilizing OSNs. These activities are resulted towards the generation of big data. Thus, such a large data repositories invite the scholastic communities to extract valuable information with the purpose of converting the contents into knowledge. As a bitsy part of research community, we accept this invitation and divert our focus towards analysis of citation networks.

In OSNs, many human behaviors have been studied. Among all of studies, social influence is a phenomenon which has been studied most from last decade. In simple terms, it reflects the behavioral change of individuals affected by others in the social network. Social influence has been studied in different aspects and has historical root in sociology. Through researching opinion formation and diffusion of influence, the role of social influence models is to extract the impact of community leader that make it possible for their followers to change behaviors. Recently, social influence has begun to attract more attention according to the available of many datasets. For

example, computer scientists design a model of academic influence to support the impact of each paper or author make in each field. This model is also used for citation network.

Citation network is a scholastic network that records the citation information of each research content in it. More precisely, if paper A cites paper B, we can show this phenomenon by incorporating an arrow heading from the node representing paper A to the node that represents paper B. Citation analysis is actually the examination of the frequency, patterns, and graph of citation in articles and books [1]. It uses the citation relationship to represent the connection with other works.

1.1 Problem Statement

The main purpose of the underlying project is to mathematically formulate the problem of ranking the communities while considering their influence on each other. Identifying the most influential community in a specific citation network is the main contribution of the project. After tireless efforts, we are now able to present the joint weight based on aggregation citation as well as the modified PageRank algorithm. In the prescribed problem, the citation network is described as a joint weight based directed graph. The nodes of the graph are representing to a particular community while the citation relationship is described as the edge of the graph. For validation of our proposed work, we apply the weighting formula to the graph in order to get the value of each edge and subsequently implement our modified algorithm for figuring out the most influential community.

1.2 Contribution

The first contribution of the project is the consideration of community-level influence in citation network. The next contribution is the novel idea of the intra and inter influence effect of different communities on each other. Within the underlying citation, papers' information also includes the venue where it has published in. For our scenario, these venues are considered as the communities. The third contribution is the consideration of internal and external citation weights of each community. Further, the level of each citing community is also considered. Finally, we validate the performance of proposed work experimentally by employing modified PageRank algorithm.

1.3 Report Structure

The report is organized as follows. We introduce the history and related work of the citation analysis briefly in Chapter 2. Moreover, the introduction of PageRank algorithm is also stated in Chapter 2.

In Chapter 3, we describe the model of underlying issue in detail and the formula of weight calculation is also presented. The modified PageRank algorithm is also included in this Chapter. Chapter 4 displays the result of the experiment and discusses the factors that affect the influence of communities. The comparative study is also presented in this part. Chapter 5 discusses and concludes all of our findings.

Chapter 2: Background

2.1 Related Work

Society has been organizing itself into communities, which are groups of individuals with common interests or topics. Brino *et.al [1]*analysed the social influence in community level. Then they studied the core impact on community underlying structure. They found that members of the community core work as bridges that connect smaller clustered research groups as well as increase the average edges degree of the community underlying network, but decrease the overall network assortativeness. And the most influential leader of the community is identified successfully. It considers the internal influence of each leader, but does not consider the external influence of each leader of the specific community. The influence of each leader on other communities has not been discussed. Jie Tang *et.al [9]* studied a novel problem of topic-based social influence analysis. They supposed a novel approach to describe the problem using a graphical probabilistic model. And they also designed a new algorithm for training the model. The experiments show that the discovered topic-based influences by the proposed approach can improve the performance of expert finding. The contribution of this paper is analysing social influence in topic level. Rong-Hua Li *et.al* [13]studied, for the first time, the influential community search problem in large networks. They present a new community model called *K-influential* community based on *K-core*. A linear-time online search algorithm and an optimal index-based algorithm are proposed. They apply their algorithms to the proposed model to generate the influential community in social network. However, the drawback of this paper is that the number of communities implemented in the experiment part is small. And the structure used in experiment is undirected.

Citation network is the emerging topic to be considered by the research community during recent era. Many nuanced results have been reported in the underlying field. Nan Ma *et.al* [12]provided an alternative method to analyse the citation network. It uses the well-known Google algorithm – PageRank algorithm to analyse the citation network. It attempts to offer a more integrated graph of citation network in a specific field. The PageRank algorithm provides a comprehensive aspect to analyse the citation network instead of using number of citation. A meaningful of advantage of PageRank they found is that it could largely eliminate the flattery of academic influence caused by author self-citations. They just used the original version of the PageRank algorithm. But the citation

relationship is not same as hyperlink in the Internet. Jiang Li *et.al* [11] suggested an alternative to the widely used Times Cited criterion for analysing citation networks. They proposed a novel algorithm called ArticleRank to provide a different ranking of a set of papers form the citation networks. This algorithm describes a modification of the PageRank algorithm. However, it requires substantial computation if the algorithm is to run for many iterations on large numbers of papers. Erjia Yan *et.al* [7] provides an alternative perspective for measuring author impact by applying PageRank algorithm to a co-authorship networks. The weighted PageRank algorithm combines citation and co-authorship network topology in a very effective way. Compared to other related PageRank, it focuses on the random surfing aspect and develops it into citation ratio. These two papers attempt to apply the PageRank algorithm to analyse the academic impact of author and article. The PageRank algorithm provides a meaningful extension to the traditionally used citation for authors. These two papers also provide some modifications of the PageRank algorithm. The method of these two papers inspire us to apply the algorithm at the community level. The algorithm we applied on the dataset is also a modified version of PageRank.

In previous works, few considered the community-level influence. What they focused on is the impact of leader in a specific community or in a specific social network. We consider the community influence analysis based on the external and internal citation weight as well as their aggregated joint weight. Further, the level of particular communities is also considered in this work.

2.2 PageRank

The PageRank algorithm is designed by Larry Page and Sergey Brin in 1998. It was presented and published at the Seventh International World Wide Web Conference (WWW) in April 1998. PageRank algorithm aims to measure the importance of each website pages using hyperlinks among pages [12]. The importance of each page is “voted” by all the other pages in the Web. A link to a page counts as a vote to support. The motivation behind the PageRank algorithm was the track some difficulties with the content-based ranking algorithms of early search engines which used text documents for web pages to retrieve the information with no explicit relationship of link present between them.

The core equation of the PageRank algorithm is presented as follow:

$$PR[A] = \frac{1-d}{N} + d(\sum_{j=1}^{n}(\frac{PR[C_J]}{[O_{C_J}]})) \qquad (1)$$

where N is the total number of pages in web. $PR[A]$ denotes the PageRank value of a particular page A; d is the damping factor, and it is set to 0.85 in the standard PageRank algorithm. C_J is one among *n* pages that links to Page A; and O_{C_J} is the number of outgoing edges of page C_J. Thus, PR[A]'s value depends on the value of Page A's incoming websites links as well as on the outgoing pages of A. Based on the core equation of PageRank, some observations can be made:

- The value of PR[A] will increase with the increase in the number of web pages that have connection with page A.
- As the higher ranked pages link to Page A, the higher rank value will Page A get.
- When the damping factor d is set to 1, the stronger the above effects.

Therefore, the motivation of PageRank for a web page A can be described in details as follows. If page A is important, the web pages connects to it must be vital too. That is the reason why we mentioned the PageRank i.e. "The importance of the page is "voted" by all other pages in the Web" [15]. Thus, the definition of the PageRank value of each page is defined as follows

$$PR[A] = \sum_{j=1}^{n}(\frac{PR[C_J]}{[O_{C_J}]}) \qquad (2)$$

The solution of this equation is: $P^n = H^n p_0$. P^n is *n* times of user surf the webpage. p_0 is the initial rank value of each page, and in the first definition p_0 is Uniform Distribution and H^n is the source matrix. IWhen some pages linked to the non-outgoing page, all the value will be sunk into underlying page. We call this kind of page "damping page". A common way to deal with the "damping node" is to assume the surfer will start a complete new web which does not have the connection to the "damping node". To eliminate the situation, the matrix is changed to:

$S = H + ed^T / N$. where N is the total number of the pages exist in the web, e is uniform matrix.

That is the reason why the damping factor was introduced in the core equation of PageRank. The value was prompted by the anecdotal observation that an individual will follow of the order of six hyperlinks, corresponding to a leakage probability (1-d) =1/6 ≈ 0.15 [2]. The value means that user

may have 15% chance to start a new page which does not have the connection to this kind of page[3]. So the source matrix is changed to: $G=ds+(1-d)ee^T/N$. Applying this matrix to $P^n = G^n p_0$., the final definition is presented as follows:

$$PR[A] = \frac{1-d}{N} + d\left(\sum_{j=1}^{n}\left(\frac{PR[C_J]}{[O_{C_J}]}\right)\right) \qquad (3)$$

2.3 Problem Definition

In this section, the basic definition of the problem is presented. For the purposed of our work, we consider citation network as a directed graph with weightage in which nodes represent venues and edge link denotes the citation relationship between venues. The community influence analysis problem is to identify the most influential community in citation network. In citation network, each venue represents a scientific conference or journal. Therefore, each venue can be considered as a community. Thus, we do not need to do employ the community detection method. The citation network is defined as *G (N, E)*, where *N* represents the node of the graph and *E* is the edge of each node. n=|N| is the total number of the venues in the underlying dataset. We apply the proposed algorithm on this model to identify the most influential community in citation network. In figure 1, we present a simple graph of the citation network:

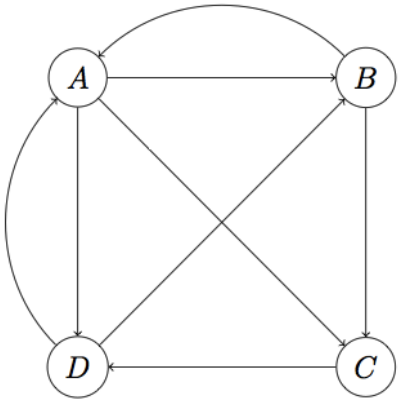

Figure 1 Simple citation network

Each node is a community, and the edge represents the citation realtionship between communities. The weight value is associated with each edge. The value of weight of each venue is calcualted by the forumula. The detail of formula is explained in detail in Chapter 3. Thus the citation network is defined as the directed graph with weightate. The most influential community is defined to consider

internal and external influenece according to weight formula. The rank value of each node is calculated by the proposed algorithm. The most influential community (node) is the node whose rank value is highest.

Chapter 3: Design and Implementation

Based on the definitions and assumptions above, the model is proposed to formula the community influence analysis problem. The algorithm implementation is also discussed in this section. We illustrated the proposed technique with pseudo code.

3.1 Design

The proposed algorithm is based on the PageRank algorithm. The proposed algorithm called ComRank. The modified PageRank algorithm applied on the refined dataset. It identified the most influential communities. However, we found that PageRank algorithm has some drawbacks when we applied it to the citation analysis problem. For example, PageRank algorithm considers external influence between each pages via hyperlinks among them. The structure used in citation analysis is directed without weightage value. However, in community level influence analysis problem, we described the citation network as directed graph with weightage. PageRank algorithm is applied on directed graph without weightage and it ignores the internal influence. Because the hyperlinks do not point to itself. So the proposed algorithm must meet the model whose structure is weightage and consider internal and external influence. The weight equation successfully meets these tasks. It considered internal and external influence of each node in the graph. The main objective of the algorithm is to identify the most influential nodes, which refers the community in the citation network.

3.1.1 Joint Weight

A new idea to study the key problem of ranking the scientific communities, based on joint weights in citation network is presented. First, we explained the concept of joint weights based on citations that a venue acquires internally as well as from external venues. The citation network in this problem is described as a directed graph with weightage value. The nodes represent community in the graph. The edge of each link between nodes can be generated by the weight equation shown as follow:

$$W_i = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{C_{ij}}{(T_i + T_j)/2} * C_{i,id} \right) \qquad (4)$$

Where C_{ij} represents that *i* cited by *j* venue, $C_{i,id}$ indicated the in-degree of cited venue *i* and T_i+T_j total citations received by both connected venues in citation network. The denominator factor is taken as a summation of both connected venues to eliminate the effect of low value if it is recorded

during citation analysis in citation network. The Figure 2 is shown below to explain in detail about the weight equation:

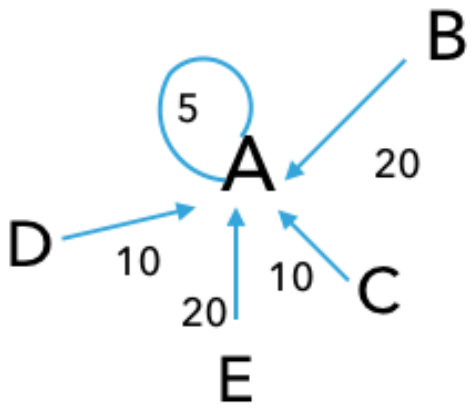

Figure 2 Weight example

In Figure 2, community A got citation form A itself, which is called internal influence, community B, C, D and E. Total citation of community A is 20+20+10+10+2=65. When calculating the weight of A from node B, assume the total citation of community B is 20. So the result of $(T_i + T_{j)}/2$ will be (65+21)/2=43, C_{ij} is 20, and $C_{i,id}$ is 5. The weightage value of A from node B is (20/43) *5.

When the graph is settled, the weight matrix can be generated by the structure of the graph. The connected node is defined as the citation relationship nodes. We can use the edge with the weight value to form the weight matrix to represent the citation relationship in citation network. The weight matrix does not only represent the connection of each node but also reflect the citation relationship among nodes. In the matrix, the rows and columns are the same as the number of connected nodes in the graph.

3.1.2 Computing ComRank

The weight matrix can be generated via the weight equation mentioned above. Weight matrix is a square matrix with rows and columns corresponding to citation relationship. If there is an edge from node *u* to *v* whose weight value is zero. Thus, some vector over the citation relationship that corresponds to a source of rank will be applied to the weight matrix. Note that weight matrix is all positive, the added vector must be reduced to balance the formula. The equation is shown as follows:

$$w' = d * w + (1 - d) * \frac{eT}{n} \qquad (5)$$

where *w* is the weight matrix, and $\frac{\mathrm{eT}}{\mathrm{n}}$ is the uniform matrix divided by the number of nodes *n.* note that d<1 is same as the damping factor as PageRank algorithm. The community rank value for a signal node is defined as follow:

$$R[u] = \sum_{j=1}^{n} \left(\frac{R[C_J]}{[O_g]}\right) \qquad (6)$$

where $R[u]$ means that the ComRank value of node u, C_J is the rank value of connected node *j* to node *u,* and O_g is the number of outgoing edges of node *j*.

During calculation, the dimension of the rank vector is (n, 1). The initial rank value of each node is set to 1/n. The modified weight matrix w' is generated by the matrix *w* of citation relationships. The purpose of the adding part e^T/n of this formula is to erase the column's vector whose value is zero, which means that the node does not have any connection to others. If not, the definition of the rank value of node is set to a temporary value. Power method is applied to calculate rank and compare the difference between present iteration's result and last iteration's result. We considered the value convergence when the difference between them is small enough.

3.1.3 Algorithm Outlines

Based on the ideas mentioned above, we proposed the algorithm as follows:

Algorithm: ComRank

Input: weight matrix *w* and *G(V,E)*

Output: Rank vector *R*;

---

1. set *threshold value*
2. set damping factor *d;*//between 0 to 1
3. *n=|G|*
4. uniform matrix e^T
5. **for** *i=0* to *n-1*
6. initial value $\rightarrow$*CR[i];*
7. transition matrix w';
8. *R[0]=w'*CR[0]*
9. **for** *i=1* to *n*
10. **while** *|R[i]-R[i-1] | > threshold value* **do**
11. *R[i] =R[i-1] * CR[i]* ;
12. **continue**;
13. **if** *|R[i]-R[i-1] |< threshold value* **then**
14. **break**;
15. **collections.sort** vector *R.*

Step1: Threshold value is used to compare the difference between present iteration result and last iteration result.

Step2: Set the damping factor. The purpose of the damping factor is to set the probability of the user in the web to random surf the complete new website. The damping factor used in this algorithm is same as PageRank algorithm to compare the result obtained by ComRank and PageRank.

Step3 to 7: Set the initial rank value of each node. Load the weight matrix and transform the weight matrix to generate matrix w'.

Step8: Calculate and noted the result of first iteration.

Step9-14: Implement iteration method.

Step 15: Calculate ComRank vector.

3.2 Implementation

3.2.1 Dataset

The dataset used in this project is taken form publically available data bank known as ArnetMiner Citation Network[4][10]. It consists of publications and citation information of scholastic content. The dataset is comprised of 2,244,021 papers with 4,354,534 citation relationships from Computer Science field. We explored the data in order to extract the information according to citation relationship and the number of citation. Moreover, the normalized dataset consists of index of paper, authors' names, title of paper, venue, abstract and citation information. And the information in the dataset is illustrated as follows: PaperTitle(#*), Authors(#@), year(#t), Publication venue(#c), index id(#index), the references of the paper(#%), and abstract(#!). An entity is defined as the information from #* to #!. We extracted venue(conference) information ranging from 2008 to 2012 i.e. the entities whose "time" is ranging from 2008-2012.

3.2.2 Citation Relation

The citation relationships are extracted as follows: if paper A gets citation form paper B, we extract the conferences of paper A and B. The conferences are marked as conference A and conference B[6]. So the citation relationship is mapped as conference A gets citation from conference B. The conferences in our problem is defined as the communities. So we can form the connection of two communities via the citation relationship between them. Among entities in the dataset, we mapped each entity's publication venue and index number. The reference information is extracted to get the referenced venue. For example:

#*Reliability analysis of waste clean-up manipulator using genetic algorithms and fuzzy methodology.

#@Naveen Kumar,Jin-Hwan Borm,Ajay Kumar

#t2012

#cComputers & OR (we call this venue C1)

#index3063624

#%812528

#%814449

#!

In this example, the index number of referenced paper is already mapped to their publication venue. We take C2 and C3 to represent the publication venue of papers whose index number is 815258 and 14449. So the citation relationship is defined in this study: C2 got citation form C1, and C3 got citation form C1. We used Java to implement this task.

The following table is the example of the citation relationship generated in experiment:

Table 1: Citation Relation

AAAI	AAAI
AAAI	AAMAS
AAAI	ACL
AAAI	ARES
AAAI	CHI
AAAI	CoPR

In the table shown above, the relationship is defined as publication venue AAAI got citation from 6 venues, which are AAAI itself, AAMAS, ACL, ARES, CHI and CoPR.

3.2.3 Weight Matrix

Based on the citation relationship generated above, we apply the weight equation mentioned in 3.1.1 to calculate the citation weight of each citation relationship. We use Java language to record the number of referenced publication venue. And the following table shows the result:

Table 2: Citation weight

Community	AAAI
AAAI	0.0519078

AAMAS	0.0639103
ACL	0.085004
ARES	0.084193
CHI	0.071380
CoPR	0.014095

The table of the citation weight is generated by the citation relationship mentioned above. The internal influence is defined as the a publish venue get citation from itself.

3.2.4 Algorithm Implementation

The detail of algorithm implementation discussed in this section. The above-mentation techniques are implemented using Java language. The edge is stored as the citation relationship in graph. The citation relationship is stored in a structure map called ConferencedNum. The weight value of each node calculated via the weight equation mentioned above is also stored in a structure map called matrix. The weight matrix is the input of the algorithm. The rank value is calculated iteratively. The following pseudo code of proposed algorithm is as follow:

```
//load graph;
Treemap ConferencedNum<…>
Treemap matrix<…>
ConnectedCom=entry.getKey();
totalCitation=entry.geyValue();
printmatrix();//generate citation relationship
/*read matrix*/
class Weight{
String C1;
String C2;
```

```
/*algorithm calculation*/
/*initization*/
damping facto d;
threshold value;
initial rank vector Ir;
Matrix ee; //uniform matrix divided by n
            //n=|G|
Matrix transition = w.times(d).plus(ee.times(1-d));
Matrix Rank= ransition.times(w).times(x);
```

```
return"C1:"+C1+"---C2:"+C2+"-
weight:"+weight;
// form the type of the citation relationship
}
/*get connected venues*/
ArrayList<String> getMetadata; //md
/*get weight information*/
ArrayList<Weight> getWeight; //wd
String [] st
st[0]=C1;
st[1]=C2;
st[2]=weight;
/*form the weight matrix*/
double [][]wm;//weight matrix
for (int i=0;i<wd.size();i++){
for (int j=0;j<md.size();j++){
if (wd.get(i).d1.equals(md.get(j))){
for (int k=0;k<md.size();k++){
if (wd.get(i).d2.equals(md.get(k))){
wm[j][k]=wd.get(i).weight
```

```
while(true){

if(compareAbs( threshold,x.minus(Rank))){
  break;
}else{
x = r;
r =transition.times ( Ir );
iter ++;}
}
rankSort();
```

The flow chart of the experiment is show as follow:

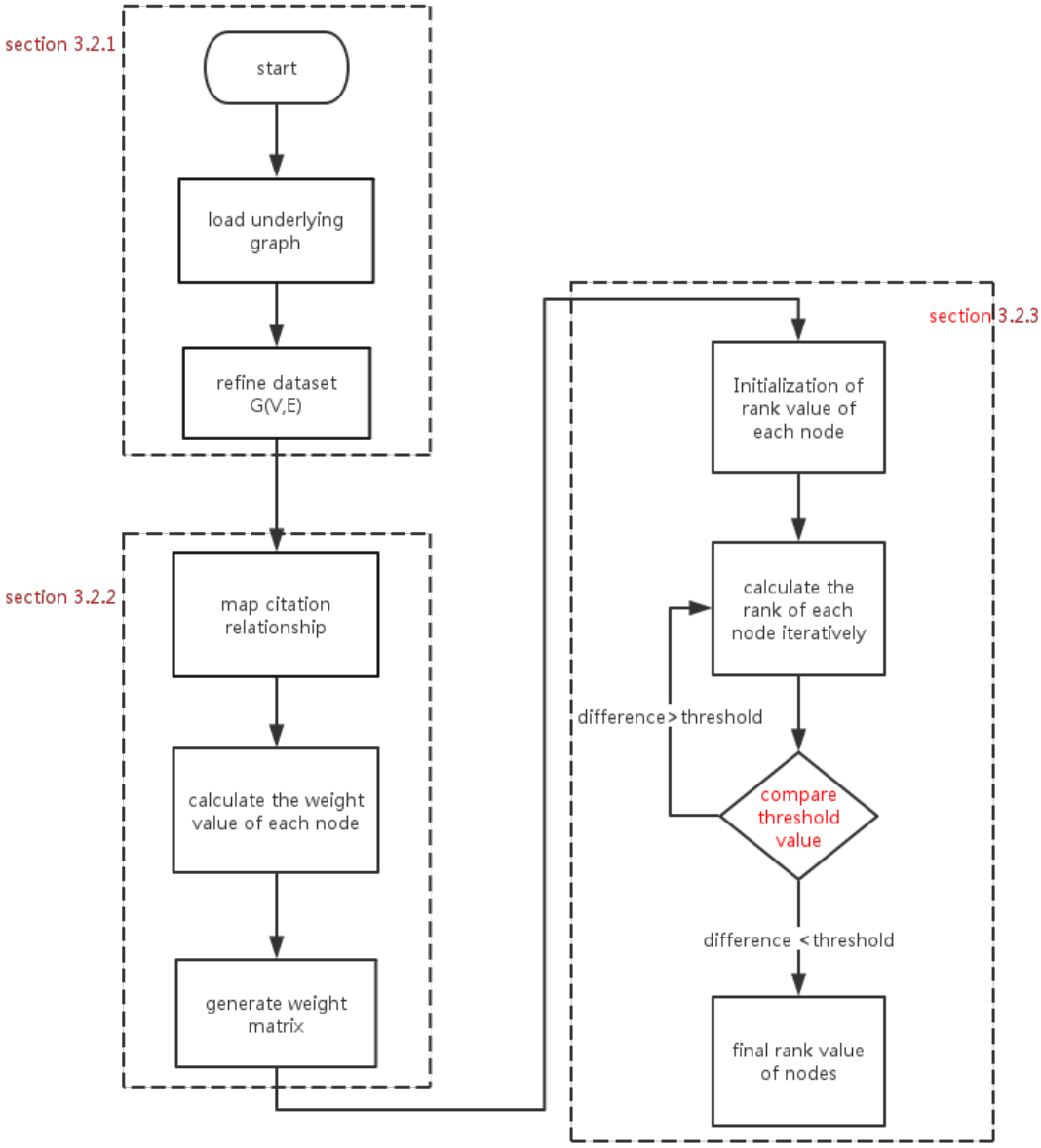

Figure 3 Flow chart of experiment

Chapter 4: Results and Discussion

The ComRank by incorporating joint weight is mentation in equation (4). After applying the proposed method to the citation network, the rank o influential venue is obtained. Many iterations were performed to produce ranks through iterative process based on the PageRank concept. The top ranked venues with maximum level through the presented novel method is of greatest interest. A series of comparative experiments to validate the results of ComRank algorithm are proposed. The rank vector considers the conference whose level information is publicly available[5]. The level of the conference is also a standard to evaluate the performance of the algorithm. The result of PageRank algorithm is obtained to compare ComRank. The performance evaluation of ComRank algorithm is proposed in different ranges. We compare the rank of top5, top 10 and top15 communities' levels of ComRank and PageRank. The maximum rank values with variant damping factors are obtained by the both algorithms. The stability comparison is also mentioned. A scatter plot will be shown to analyze the stability of algorithm.

4.1 ComRank with different damping factors

The rank values are computed by the ComRank and PageRank as shown in table 1. The maximum rank values for different damping factors are calculated.

Table 3: Different damping factor & Max value

Damping factor	Max value(10^{-19})	
Algorithm	ComRank	PageRank
0.85	3.02	2.10
0.75	2.53	1.76
0.65	2.87	2.01
0.55	2.37	1.66
0.45	2.34	1.64

0.35	2.20	1.55
0.25	2.90	2.08
0.15	1.36	1.00

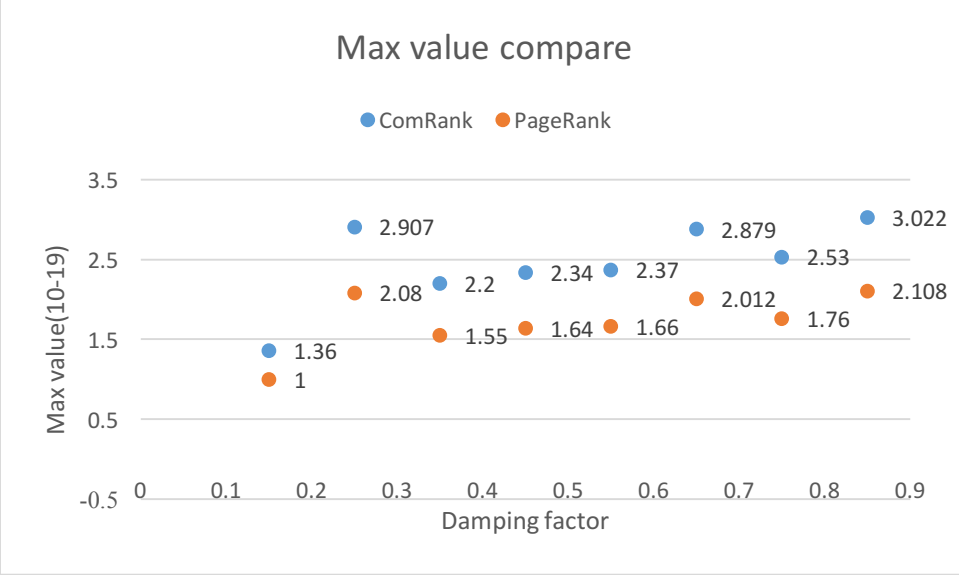

Figure 4 Damping factor & Max value

Figure 4 shows scatter plot of max vlaues generated by ComRank and PageRank for different damping factors. We can find that the max vlaue of ComRank is always above the value of PageRank. That is not suprising since PageRank algorithm is more related to external impact of community. The ComRank considers internal and external impact of each community.

4.2 Community Ranking

To evaluate the performance of the algorithms, we generated different top communities of both algorithms. The level of the community is the criterion. If the result shows more high-level community, then the given result is better. The number of high level communities is also the criterion of the evaluation. The results are given in Table 4, Figure 5 and 6 that show the clear picture of top ranked venues produced by both algorithms. We generated top 5, top 10 and top 15 communities.

4.2.1 Top 5

Table 4: Top 5 community

Rank	ComRank		PageRank	
NO.	Community	Level	Community	Level
1	AAAI	A	SAC	C
2	ICDE	A	ICML	A
3	SIGCOMM	A	IJCAI	A
4	CRYPTO	A	GECCO	B
5	KDD	A	AAAI	A

The Table 4 shows the top 5 venues produced by PageRank and ComRank. We achieved better results than the PageRank where low level conferences are produced to be a high rank conference, for example SAC is a level C conference but the PageRank predicted it as a topped ranked conference. Similarly, PageRank detected some level B conferences as topped ranked. On the other hand, ComRank detected only topped ranked conferences. The first place community of PageRank algorithm is SAC community. The ComRank algorithm generated AAAI at the first place, which is a very famous academic conference around the world. In this way, the performance of the ComRank algorithm is better.

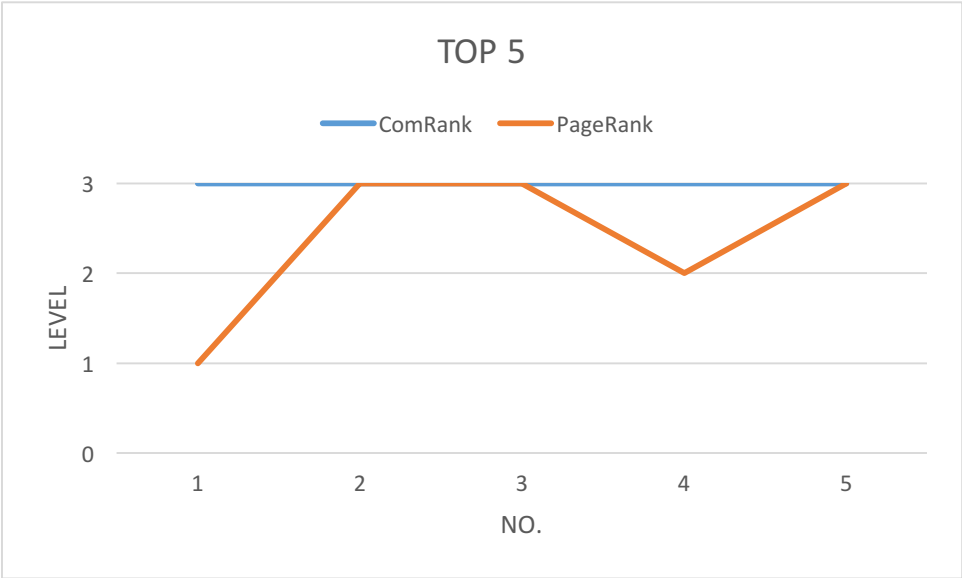

Figure 5 Top 5 - line chart

1---- C level 2----B level 3-----A level

The stability of ComRank and PageRank is shown in Figure 5. ComRank generated stable result. It detected only topped ranked conferences. The stability of PageRank algorithm is worse than ComRank algorithm.

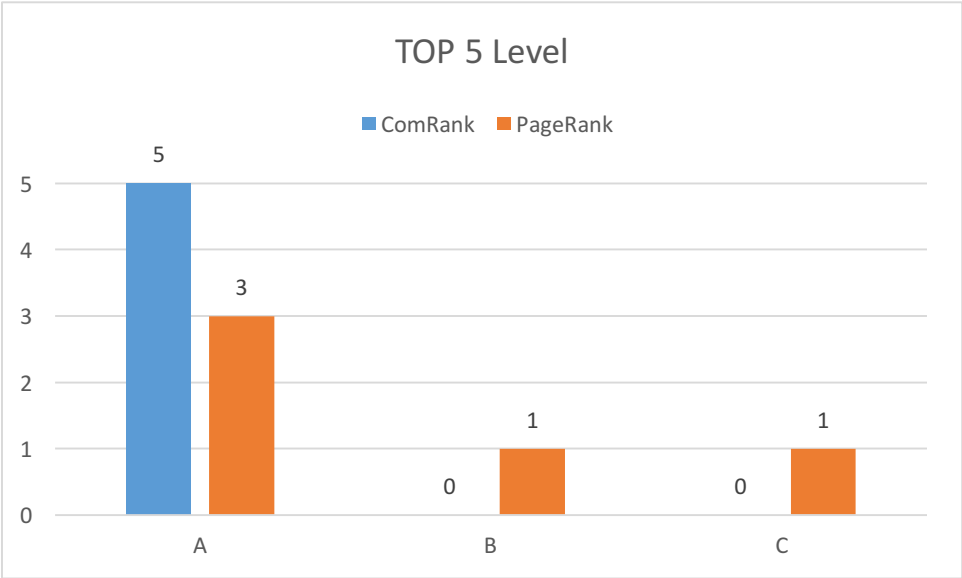

Figure 6 Top 5 level

The number of different level communities is shown in Figure 6. The number of A level communities of PageRank algorithm is 5. The number of high level communities existing in the result is the sign of better performance of algorithm. As we take the community level to be the criterion of the

performance. High level community shown in result means better performance.

4.2.2 Top 10

Table 5: Top 10 community

Rank	ComRank		PageRank	
NO.	Community	Level	Community	Level
1	AAAI	A	SAC	C
2	ICDE	A	ICML	A
3	SIGCOMM	A	IJCAI	A
4	CRYPTO	A	GECCO	B
5	KDD	A	AAAI	A
6	SODA	A	SODA	A
7	SIGIR	A	WWW	A
8	CIKM	A	SIGCOMM	A
9	ICCV	A	EMNLP	B
10	ICCAD	A	STOC	A

Table 5 shows the top 10 venues produced by PageRank and ComRank. We achieved much better result than PageRank where low level conferences are detected. For example, EMNLP is a B level conference but PageRank still predicted it as topped ranked conference. ComRank detected only topped ranked conferences. In this way, the performance of the ComRank is better.

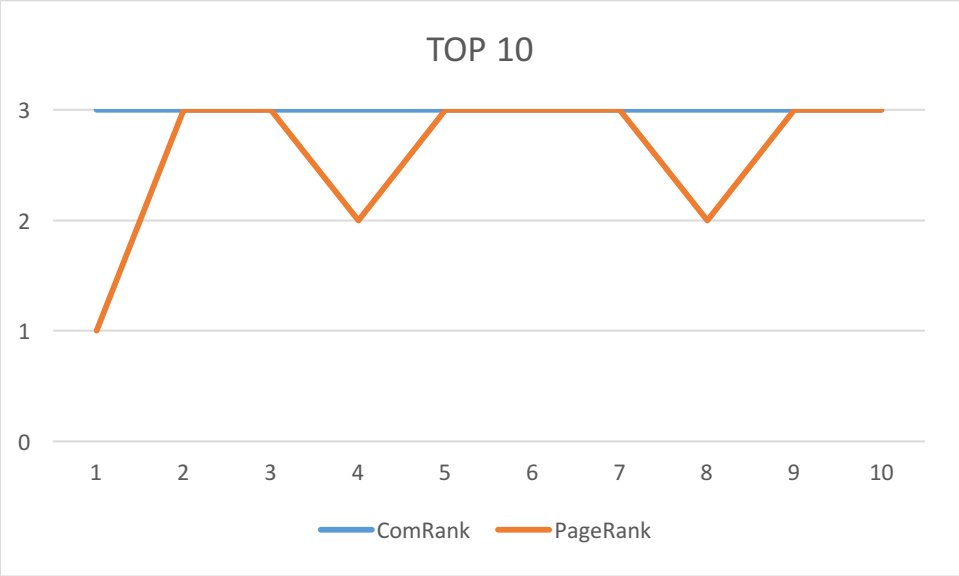

Figure 7 Top 10 wave chart

1---- C level 2----B level 3-----A level

Figure 7 is the top10 communities wave chart. The ComRank algorithm is still stable, as the 10 communities is A level. PageRank algorithm is fluctuant.

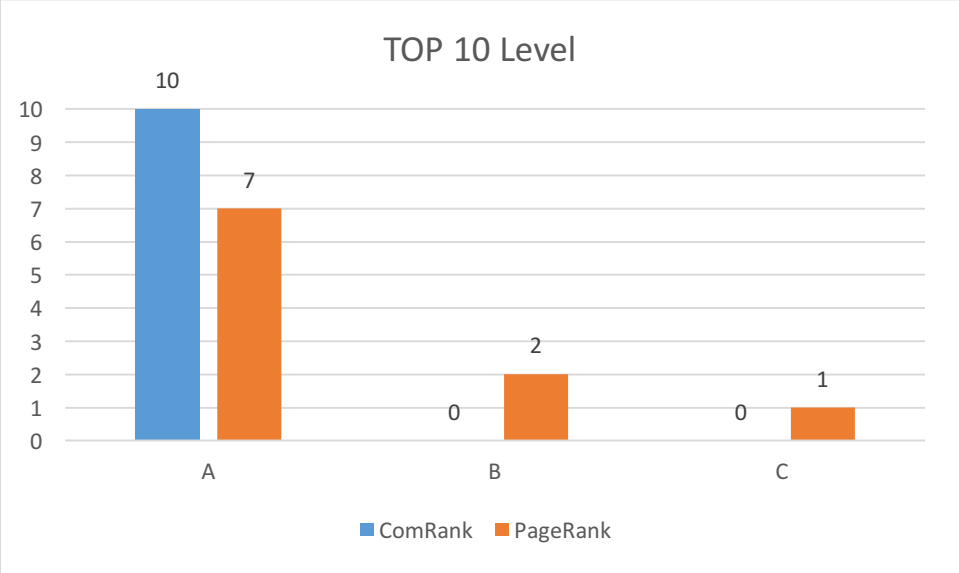

Figure 8 Top 10 level

Figure 8 shows the number of different level community exist in top10 communities. Compare the number of the high level community, the number of A level community in ComRank is bigger than that in PageRank algorithm. And about the number of low level community in both algorithms, the number of B and C level community in PageRank is bigger. The number of A level communities in

ComRank algorithm is 10. And in PageRank algorithm is 7. Not only the rank of ComRank is better than PageRank, but also the stability is better.

4.2.3 Top 15

Table 6: Top 15 community

Rank	ComRank		PageRank	
NO.	Community	Level	Community	Level
1	AAAI	A	SAC	C
2	ICDE	A	ICML	A
3	SIGCOMM	A	IJCAI	A
4	CRYPTO	A	GECCO	B
5	KDD	A	AAAI	A
6	SODA	A	SODA	A
7	SIGIR	A	WWW	A
8	CIKM	A	SIGCOMM	A
9	ICCV	A	EMNLP	B
10	ICCAD	A	STOC	A
11	TACAS	B	CRYPTO	A
12	IJCAI	A	KDD	A
13	ICML	A	SIGIR	A
14	FOCS	A	ICDE	A

Table 6 show the top 15 venues produced by PageRank and ComRank. We achieved much results

than PageRank where low level conferences are detected to be a high level conference. However, some low level conferences detected in the result, the number of low level conferences of ComRank is smaller than that of PageRank. In this way, ComRank outperforms the PageRank.

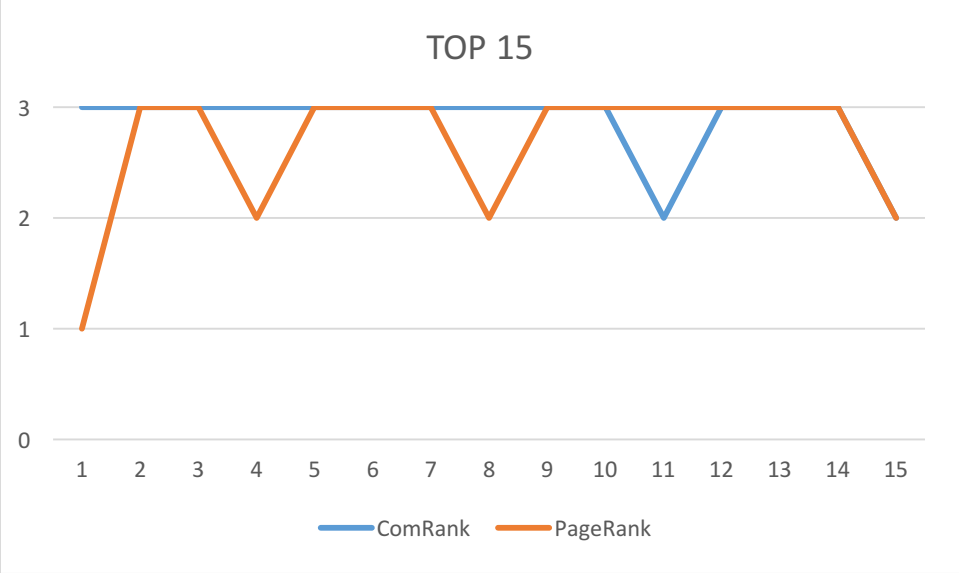

Figure 9 Top 15 Line chart

1---- C level 2----B level 3-----A level

The stability of top15 communities in ComRank is worse than top5 and top10, as two B level communities shown in the result. There is some fluctuation in ComRank algorithm. Overall, the stability is sill better than PageRank.

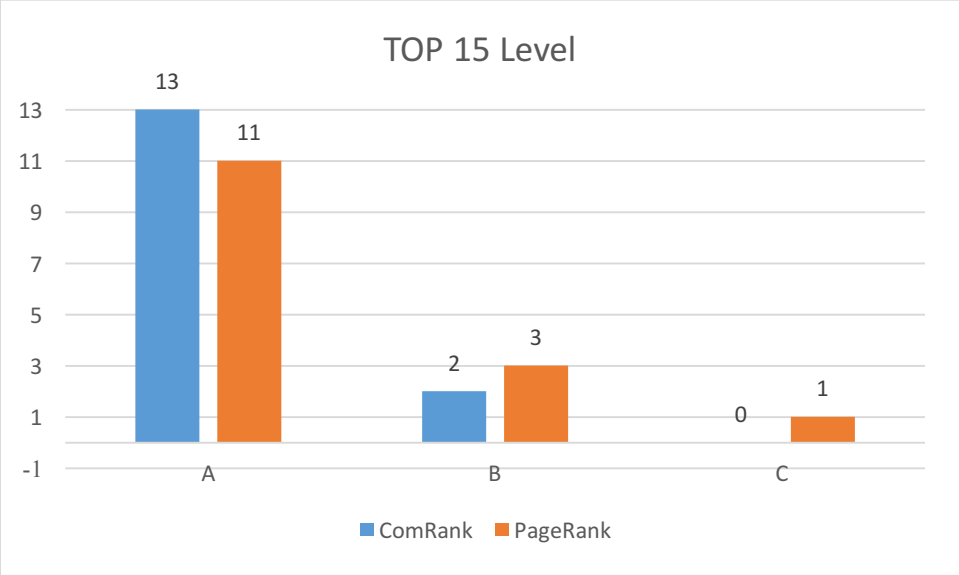

Figure 10 Top 15 level

The number of different level communities in top15 is shown in Figure 10. The number of high level communities of ComRank is still more than PageRank.

Based on different rank results and max value, we found that the ComRank algorithm is more stable and the maximum value is bigger than PageRank algorithm's. About the number of high-level communities, ComRank also has better performance. According to final result of ComRank algorithm. The first C level community shows up in the twentieth place. However, the C level community is topped ranked in PageRank algorithm. Clearly, in this way, ComRank algorithm outperforms PageRank algorithm in citation analysis. The conference level information available online is limit, we only show the result of top 15 ranking communities. the number of conference information available is about 200 conferences.

Chapter 5: Conclusion and Further Work

5.1 Conclusion

In this report, a mathematical definition of community influence analysis is proposed. We have described a modification of the PageRank algorithm that can be used to analyze the citation network. The present study attempts to measure the importance of venues from another perspective. The joint weight is presented in citation network where degree of cited venues is also considered. The proposed weights study the effect of both internal and external citations within the network. The algorithm, called ComRank, successfully identifies the most influential community in citation network. We compared the maximum values produced by ComRank and PageRank and the top ranked venues. The topped ranked venues generated by PageRank included low level communities. However, ComRank detected only topped ranked communities. The number of low-level communities produced by ComRank algorithm is smaller than PageRank. In this way, ComRank outperforms PageRank.

5.2 Further Work

Two areas for further work suggest us. First, the present experiment only considers the conference level available. However, there are many journals exist in the dataset too. We do not implement the algorithm on journal. In further work, the experiment can will consider the journal information to identify the most influential journal in citation network. Second, a drawback of the algorithm is the community level should be known. If the level of a community is obscure to us. We can not judge the rank value to identify the influence of the community. In further work of the project, we can combine the citation count to the joint weight calculation to analyze the impact of community integrally.

References

[1]. Bruno Leite Alves et.al (2013) Role of Research leader on the Evolution Scientific Communities, in *WWW 2013 Companion, May13-17,2013 Rio de Janeiro, Brazil. ACM 878-1-4503-2-38-2/13/05*

[2]. Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). *Finding scientific gems with Google's PageRank algorithm*. Journal of Informetrics, 1, 8–15.

[3]. Changhai Lu, (2011) *, mathematics behind Google, Mathematics Culture Feburuary,2011*

[4]. Citation Network Dataset, available form: https://aminer.org/billboard/citation

[5]. Conference level, available form: http://www.ntu.edu.sg/home/assourav/crank.htm.

[6]. Cawkell, A. E. (1971) Science Citation Index: Effectiveness in locating articles in the anaesthetics field: 'Perturbation of ion transport', *in British Journal of Anaesthesia, 43: 814*,

[7]. Erjia Yan, Ying Ding, (2010) Discovering author impact: A PageRank perspective, in *Information Processing and Management 47(2011) 125-134*

[8]. Graham, Paul (November 2005). "Web 2.0". Retrieved 2006-08-02.

[9]. Jie Tang *et.al* (2009), Social Influence Analysis in Large-scale Networks*, in KDD'09 June28-July1,2009, Paris,France.*

[10]. Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. (2008) ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (SIGKDD'2008). pp.990-998.

[11]. Li, J., Willett, P. (2009) *ArticleRank: A PageRank-based alternative to numbers of citations for analysing citation networks,* 61(6), pp. 605-618

[12]. Nan Ma, Jiancheng Guan, Yi Zhao (2008), Bring PageRank to the citation analysis, *in Information Processing and Management 44(2008) 800-810* Pawan Lingras, Saint Mary, *Building an intelligent Web: Theory and Practice*

[13]. Pawan Lingras, Saint Mary, *Building an intelligent Web: Theory and Practice*

[14]. Rong-Hua Li *et.al (*2014) Influential community search in large network，*in VLDB Endowment, vol.8, No.5*

[15]. Shatakirti, *Hyperlink based search algorithms -PageRank and HITS*