Finding Influential Papers in Citation Networks

Sheng Zhang*, Danling Zhao†, Ran Cheng‡ Jiajun Cheng* Hui Wang*
College of Information Systems and Management
National University of Defense Technology,
Changsha, P. R. China
*Email: {zhangsheng,jiajun.cheng,huiwang}@nudt.edu.cn
†Email: zhaodanling11@163.com
‡Email: 1585536214@qq.com

***Abstract*—The citation network is the social network made up of papers and their citation relationships. A key task of the citation network is to find influential papers in the network. Traditional properties such as centralities can not reflect the influence of papers comprehensive, since it does not take the authority of the paper and the transfer effect into account. Other algorithms such as PageRank and HITS overcome those shortcomings. However, both of them involve matrix multiplication and repeated iterative process, which is less-effective. Comparing with another often mentioned network, the coauthor networks, we noticed that the citation network is a Directed Acyclic Graph(DAG) and it has a transitive relation. Making the most of properties of the citation network, we draw on the thought of topological sorting and design a more effective algorithm, whose time complexity is linear with the number of vertices and edges, which is O(V+E). In addition, we illustrate that the algorithm we proposed is stable and effective. We also apply our algorithm to another DAG task and results show that our algorithm has great scalability.**

I. Introduction

The citation network[1] is a social network that contains papers and their citation information. In this network, the vertex stands for the paper and if there is a citation to paper B in paper A, then an edge starts from vertex A to vertex B. It is obvious that the citation network has the following properties. First, citation networks are directed, the link from one point to the other. In addition, citation networks are acyclic, because a paper only cites published papers. Therefore, the citation network is a Directed Acyclic Graph(DAG).

It is of great importance to measure the most influential paper in a citation network. A simple method is using citation times. The more times the paper is cited, the more influential the paper is. However, the great drawback is that the paper published earlier has more cited times. An improvement approach is calculating nodes' importance of a network by using PageRank algorithm[2] and HITS algorithm[3]. The idea of these methods is that a paper cited by an important influential paper may be influential. However, both of them involve matrix multiplication and repeated iterative process, which is less-effective.

Since the network satisfies the property of Directed Acyclic Graph(DAG), making the most of this property, we draw on the thought of topological sorting to design a more effective algorithm. In this model, we learn from the energy transfers in the food chain and define an Initial Contribution Coefficient to measure its authority. In addition, we define a Self Contribution Coefficient (S) to measure the influence from other papers. Finally, we design an algorithm to calculate each paper's Final Contribution Coefficient to measure the paper's influence.

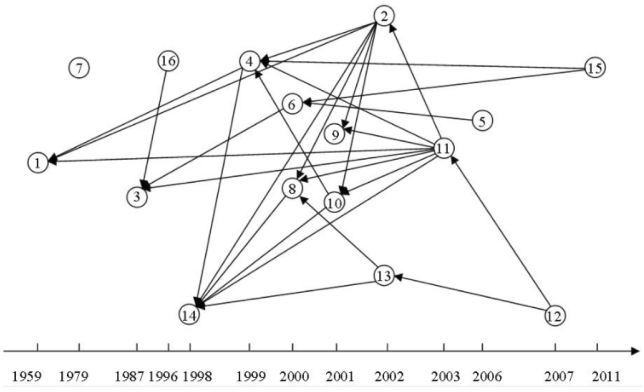

Fig. 1. A citation network consist of 16 papers.

The construct of the paper is organized as follow: In Section II, we compare similarities and differences between two similar networks, coauthor networks and citation networks. In Section III, we illustrate related works about citation networks and finding influential papers in networks. In Section IV, we designed an algorithm according to the properties of citation networks. In Section V, experiments on papers networks are illustrated and the parameter is discussed, after that we apply the algorithm to another dataset to show the scalability of our algorithm.

II. Differences to Co-author Networks

Shown in Figure 1 is a citation networks. The networks are made up of 16 papers, which are listed in chronological order. In this graph[4], a vertex reprsents a paper and the edge poingting from 11 to 1 means paper 11 cited paper1. This graph do not only reflect the citation relation, but also reflect the publication time.

Clearly, it is a Directed Acyclic Graph (DAG). Assuming that paper A cited paper B and paper B cited paper C, we can know that C can not cited A because of the publication time. That is to say, the relationship between any two papers satisfies strict partial order relation.

978-1-5090-1192-6/16 $31.00 © 2016 IEEE
DOI 10.1109/DSC.2016.55

IEEE computer society

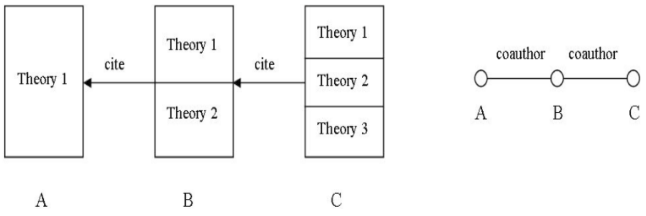

Fig. 2. The citation relationship and the coauthor relationship.

Another similar network which is always mentioned at the same time is coauthor network[5][6][7]. The coauthor networks consist of verities which represent authors of papers and edges which stand for cooperation between them. The coauthor network is a social network and it is an undirected graph.

However, the citation is a DAG because the citation relation is related to publication time. As they are both complex network, there are some similar characteristic.

In both network, the centrality can be used to measure the importance of a vertex in the network. If a researcher has more coauthor, he is more important in the coauthor network. In the same way, if a paper has been cited many times, it must be important.

However, there are still many differences between them.

In the citation network, there exists transitive relation while in the coauthor network, there does not exist transitive relation.

To be more specific, shown in Figure 2, if A cited B, B cited C, A also influenced C. But if A coauthored with B, B coauthored with C, A would not influence C. Hence, we cannot only use centrality to measure the influence of a paper. It is clear that a foundational paper in one filed may be more influential than those new papers cited many times. In addition, the publication time of a paper also affects its influence.

III. Related Works

Citation networks have been investigated by a great many researchers. Sven Bilke and Carsten Peterson[8] illustrate the topological property of citation networks. The experiment illustrates that the citation networks are tree-like in structure. Larry Page et al[2] pose PageRank algorithm to ranking web pages, and a lot of investigation[9][10][11][12] have been done in applying PageRank to rank several kinds of networks. Ying Ding et al.[13] use PageRank algorithm to rank authors in co-citation networks. They also compare PageRank and weighted PageRank results with the citation ranking, h-index, and centrality measures. However, they just rank authors instead of papers themselves. Singh et al.[14] propose an efficient method, based on citation network, to rank the research papers from various fields of research published in various conferences over the years. The importance of a research paper is measured by peer vote and then they use a modified PageRank algorithm to rank papers by authoritative score. However, PageRank involves matrix multiplication and repeated iterative process, which may be less efficient.

Therefore, in this paper, we are inspired by the topological property of citation networks, and proposed a more efficient algorithm to rank papers in citation networks.

IV. Methodology

The property of a network such as In-degree, closeness, betweenness and eigenvector, etc. In general, the times cited count, indirect citation relation, publication time and publication of a paper are related to its influence. In this part, we do not consider the effect of publication alone. The first reason is that there is no authority measure to evaluate a publication. Another reason is the times cited count also partly consider the publication because the paper published in famous publication has more chance to be cited.

Hence, it is necessary to consider improving these measures and developing a new model which makes full use of the property of the DAG.

A. Contribution Coefficients

The distance between two papers is of great significant. The shorter the distance is, the more influential between two papers. This is quite similar to the energy transfers in the food chain. In the food chain, the energy transfers in one way and it decreases progressively in the transfer process, which inspires us to build a new model.

First, we define a contribution coefficient(C) of each paper, and the initial value of those vertices without in-degree is judged by its times cited count and publication time.

$$C_i = \frac{N_{cited}(i)}{t_{now}(i) - t_{pub}(i)}, \tag{1}$$

Where: $N_{cited}(i)$ represents the times cited count of Paper i; $t_{pub}(i)$ represents the publication year of Paper and $t_{now}(i)$ represents this year.

Then, we define a Self Contribution Coefficient(S). If paper A cited paper B and is cited by C, we know that its influence in the network partly relies itself and partly relies on the reference of B. Hence, the influence of these vertices whose in-degree is not zero can be measured by:

$$C'_i = C_i + \sum_{p \in V_{in}(i)} \frac{C_p(1-S)}{d_{out}(i)}, \tag{2}$$

where $V_{in}(i) = \{k|$there is an edge points s from k to $i\}$ and $d_{out}(i)$ is the out-degree of the vertex i.

B. Algorithm

Traditionally, PageRank Algorithm and HITS Algorithm are always used to calculate the importance of nodes in a network. However, both of them involve matrix multiplication and repeated iterative process, which is less-effective. Since the network satisfies the property of Directed Acyclic Graph(DAG), we draw on the thought of topological sorting[8] to design a more effective algorithm.

After that, we design an algorithm to solve the model. The description of our algorithm is:

Step 1 Give each vertex an initial contribution coefficient according to the above formula.

Step 2 Start from the vertices whose in-degree are 0 and distribute part of their contribution coefficient equally to other vertices which they link to. Then delete them.

Step 3 Repeat Step 3 until the last vertex is deleted.

Step 4 Sort the vertices order by contribution coefficient.

And the pseudo code is shown below,

Algorithm 1 Contribution Coefficients Algorithm

Require: V vertices, E Edges, d in-degree and out-degree

function AL-CITATION(V, E, d)
  for all $i \in V$ **do**
    $C_i \leftarrow N_{cited}(i)/[t_{now}(i) - t_{pub}(i)]$
  end for
  while $d_{in}(i) = 0 \wedge d_{out}(i) \neq 0$ **do**
    for all $E_{ij} = 1$ **do**
      $C_j \leftarrow C_i(1 - S)/d_{out}(i)$
      $d_{in}(j) \leftarrow d_{in}(j) - 1$
    end for
    $C_i \leftarrow C_i \times S$
    delete vertex i
  end while
  Sort(C)
  Print(C)
end function

Since this network is Directed Acyclic Graph(DAG) and its solution process is a topological sorting process, the algorithm terminates in finite steps. In detail, there are V vertices and E edges in the network, so the time complexity of the algorithm is O(V+E). Namely, our algorithm is efficient and easy to implement.

Finally, we get each vertex's contribution coefficient and it can measure each vertex's influence in this network.

V. Experiments and Results

We collect 16 foundational papers in social networks and then we find some papers which cited more than two papers among these 16 papers and have been cited many times by means of the Web Of Science and Google Scholar. Then, the network is constructed according to their citation relationship. Figure 3 shows the structure of the network.

A. Experiment on centralities

To begin with, software UCInet is used to calculate the centrality of 16 papers in this network. The respective sequence of Degree Centrality, Closeness Centrality, Betweenness Centrality and Eigenvector Centrality are shown in Table I, which is sorted descend and numbers in cells are numbers of papers.

Obviously, the result that is measured by these centralities is inconsistent. As these measures just reflect a single measure of this network and do not treat each vertex differently. That is to say, centralities cannot reflect the influence of papers comprehensive because it does not take the paper's authority and the transfer effect into account.

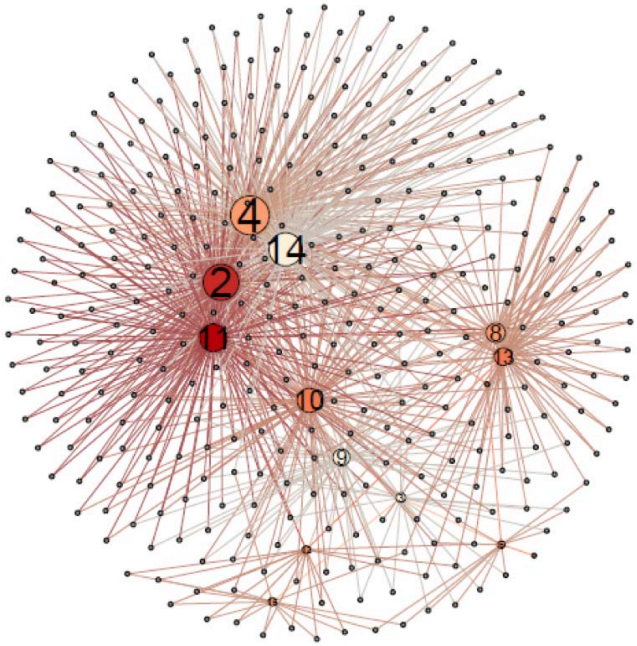

Fig. 3. Network of the papers' citation relationship.

B. Experiment on our algorithm

Now, the key problem is the value of Self Contribution Coefficient(S). Without doubt, we have no way to measure each paper's self-contribution. Hence, we regard it as 0.75 here and we will discuss the influence of S in the following part.

We calculate the contribution coefficient of 16 given papers and list in Table II.

As a result, these papers are most influential: Number 14, 4, 2. In these papers, On the one hand, Number 14 has the highest contribution coefficient. And as shown in Figure 3, paper 14 lies in the key position of the network, which connects to many other papers. Hence, the result takes the influence of

TABLE I
Four kinds of centralities.

In Degree	Closeness	Betweenness	Eigenvector
4	2	11	14
2	11	2	4
14	4	4	1
11	14	12	8
10	8	13	10
8	10	10	2
13	9	8	9
9	13	16	11
3	1	6	3
12	12	1	13
16	3	9	16
6	15	14	12
1	6	15	6
15	16	3	15
5	5	5	5

direct cited into account. On the other hand, Number 2 and 4 do not have high initial contribution coefficient, but they have been cited by some paper with high contribution coefficient. Therefore, our result shown in table II comprehensively considers both direct cited and indirect cited.

C. Stability Test

In our model, the value of self-contribution coefficient(S) may influence the result. Now, we set the value of S to be 0.5, 0.75, 0.9 and a random value (0.5-0.95) to see its influence. Table shows the sequence of the most influential paper of different value of S.

As is known from Table III, sequence of papers' contribution coefficient is quite stable. When S is given a random value from 0.5 to 0.95 in the algorithm, only 2 and 4 exchange their position in the list. That is to say, our model is reasonable and equipped with high stability.

It is obvious that the result of PageRank[13] is similar to the result calculated by our model. However, as the quantity of papers becomes large, the effectiveness of our algorithm can be highlighted. It is worthwhile for us to design a new algorithm to solve DAG problem.

In summary, our model has the following strengthens. First of all, we make full use of the properties of citation networks, namely, it is a Directed Acyclic Graph. Therefore, we draw on the thought of energy transfer in biologic chain to build our algorithm and apply it to the corporate ownership network. When comparing results we obtained with reality, we find they are quite consistent. In addition, the algorithm that we write by drawing on the thought of topological sorting is equipped with high stability and applicability. Also, the time complexity of our algorithm is O(V+E). So it is efficient and easy to realize.

In order to show the great scalability of our algorithm, we extend our model and algorithm to other data sets.

TABLE II
THE CONTRIBUTION COEFFICIENT OF 16 GIVEN PAPERS.

Paper Number(1-8)	C	Paper Number(9-16)	C
14	2037	3	479
4	1347	1	426
2	1128	16	218
11	869	6	209
10	767	12	209
9	738	7	103
8	716	5	78
13	567	15	78

TABLE III
THE VALUE OF SELF-CONTRIBUTION COEFFICIENT(C) WITH DIFFERENT VALUES OF S

$C(S=0.5)$	$C(S=0.7)$	$C(S=0.9)$	$C(S(0.5-0.9))$	PageRank
14	14	14	14	14
4	4	4	2	1
9	2	2	4	4
2	11	11	11	2
1	10	10	10	8
11	9	8	9	10
3	8	9	8	11

D. Application on other DAG problems

In this part, we choose the data of US Companies to apply our algorithm and model we used in the citation network to another DAG problem. Firstly, we construct the network according to the corporate ownership. Then we define some new metrics which are similar to those in citation networks. Finally, we analyze some influential factors to evaluate the network.

1) Construct the Ownership Network: The dataset[15][16] consists of 8,343 companies and 6,726 relationships. Since the type of companies may affect their evaluation of influence, we choose about 500 media companies to construct a network to analyze. Each vertex represents a company and an edge pointing from X to Y means company X owns company Y. Since we do not have ownership relationships for all companies, there will be companies without links.

Obviously, this is a DAG like citation network. Hence, we use the model and algorithm to identify some prominent firms.

Kim Norlen et al.[15] use two metrics to evaluate the ownership network: company degree (the number of relationships each company has) and component size (number of companies connected together). The result can be easily given by using UCInet to calculate the degree and betweenness. However, these two metrics neglect the value of a company, which is similar to the initial contribution in citation network.

In this network, we can first calculate the initial value(V) of each company. We know many factors affect the evaluation of a company, such as cash, stock, real estate, technical personnel, patent and relationships, we use the following formula to calculate the initial value:

$$V = a_1 v_c + a_2 v_s + a_3 v_{re} + a_4 v_{tp} + a_5 v_p + a_6 v_r, \qquad (3)$$

Where: $v_c, v_s, v_{re}, v_{tp}, v_p, v_r$ represent the value of case, stock, real estate, technical personnel, patent and relationships. $a_1, a_2 \cdots a_6$ represent weighting coefficient of those factors respectively.

Then we can use the proportion of stock to measure the control ability of parent company, which is corresponding to the self-contribution coefficient(S), using the formula shown as follow:

$$S = \frac{v_s(sub)}{v_s(parent)}, \qquad (4)$$

Where: S represents the control coefficient between parent company and its subsidiary; $v_s(sub)$ represents the subsidiary's stock value; $v_s(parent)$ represents the parent company's stock value.

After defining the new metrics under the new background of corporate ownership, we can start to use the algorithm to get the result. Since there are too many company and the data is not complete, we set the same initial value of companies. Finally, we will change the initial value to analyze its effect of the result.

2) Result and Analysis: The result is shown in Table IV.

In order to test the result, we find the top 20 companies ranked in 2001 from the web.

There are 9 companies list in Table V can be found in top 20. That is to say, more than half of 15 companies in our most influential list are admitted by the authority. Hence, our model and algorithm can be implemented in this data set.

To sum up, in this part, we choose about 500 US Media Companies and construct an ownership network. As both of corporate ownership network and citation network is DAG, we directly use the model to get some influential companies. Then we find some ranking information of these influential companies to verify out result. These influential companies are: Clear Channel Communications, Comcast, Gannett, Hearst, Cox Enterprises, Vivendi Universal, Viacom. Overall, the model of citation network can be used to solve some problems with directed network, especially solve those DAG problems.

VI. Conclusion

Finding a measure for citation networks is of great importance, due to the necessity of evaluation of the influence of papers. Many algorithms such as PangRank are used to do the task, but these methods involve lots of computation. In this paper, we make the most of topological property of citation networks and we draw on the thought of energy transfer in a food chain to build our algorithm. In our experiment, it is shown that our algorithm is better and more efficient than PageRank because of the simplified calculation. After that, we apply our algorithm to the data of US companies ownership networks to show the great scalability of our algorithm.

In the future, we will extend our method to larger datasets and analyze the performance of the algorithm comparing to PageRank and other algorithms.

TABLE IV
Result from the algorithm.

Name	Score	Indegree(d_{in})
Clear Channel Communications	3925	0
Liberty Group Publishing	2231.1	0
AT&T	1866.06	1
CNHI	1577.86	0
Comcast	1553.31	2
Liberty Media	1347.1	2
Media Central	1119.87	0
Disney	1116.73	0
Gannett	1059.87	0
Lee Enterprises	970.924	0
Bertelsmann	947.431	0
Viacom	875.479	2
Vivendi Universal	755.789	0
Hearst	690.283	0
Cox Enterprises	623.27	0

TABLE V
Top 20 companies ranked in 2001 from the web.

Rank	Company	Rank	Company
1	Time Warner Inc	11	NBC Universal Media
2	ViVendi Universal	12	Tribune
3	The Walt Disney Co	13	McGraw Hill
4	Viacom Inc	14	Cablevision
5	Comcast	15	Charter
6	Sony	16	Hearst
7	News	17	EchoStar
8	Cox	18	Adelphia
9	Clear Channel	19	The New York Times
10	Gannett	20	The Washington Post

Acknowledgment

The authors would like to thank the university, National University of Defense Technology, for providing a comfortable working atmosphere. The research is supported by National Natural Science Foundation of China (No.71331008) and (No.61105124).

References

[1] M. Newman, *Networks: an introduction*. OUP Oxford, 2010.
[2] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web." 1999.
[3] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.
[4] D. Yaru, "Structure modeling of citation network systems [j]," *Library and Information Service*, vol. 4, pp. 58–61, 1996.
[5] D. Zhonghua, "The comparison among social network, citation network and link network [j]," *Library Journal*, vol. 9, p. 003, 2008.
[6] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 807–816.
[7] Y. H. Said, E. J. Wegman, W. K. Sharabati, and J. T. Rigsby, "Retracted: Social networks of author–coauthor relationships," *Computational Statistics & Data Analysis*, vol. 52, no. 4, pp. 2177–2184, 2008.
[8] S. Bilke and C. Peterson, "Topological properties of citation and metabolic networks," *Physical Review E*, vol. 64, no. 3, p. 036106, 2001.
[9] S. White and P. Smyth, "Algorithms for estimating relative importance in networks," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 266–275.
[10] T.-Y. Liu, "Learning to rank for information retrieval," *Foundations and Trends in Information Retrieval*, vol. 3, no. 3, pp. 225–331, 2009.
[11] Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum, "An empirical study on learning to rank of tweets," in *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 295–303.
[12] X. Cheng, S. Su, Z. Zhang, H. Wang, F. Yang, Y. Luo, and J. Wang, "Virtual network embedding through topology-aware node ranking," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 2, pp. 38–47, 2011.
[13] Y. Ding, E. Yan, A. Frazho, and J. Caverlee, "Pagerank for ranking authors in co-citation networks," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 11, pp. 2229–2243, 2009.
[14] A. Singh, K. Shubhankar, and V. Pudi, "An efficient algorithm for ranking research papers based on citation network," in *Data Mining and Optimization (DMO), 2011 3rd Conference on*. IEEE, 2011, pp. 88–95.
[15] K. Norlen, G. Lucas, M. Gebbie, and J. Chuang, "Eva: Extraction, visualization and analysis of the telecommunications and media ownership network," in *Proceedings of International Telecommunications Society 14th Biennial Conference (ITS2002), Seoul Korea*. Citeseer, 2002.
[16] M. Gebbie, K. Norlen, G. Lucas, and J. Chuang, "Improving transparency: extracting, visualising and analysing corporate relationships from sec 10-k documents," *International Journal of Technology, Policy and Management*, vol. 7, no. 1, pp. 15–31, 2007.