Application of network analysis methods in the selection of information sources

I G Olgina

Omsk State Technical University, pr. Mira 11, 644050, Omsk, Russia

E-mail: inna_olgina@mail.ru

Abstract. A possibility of applying network analysis for solving practical tasks related to the selection of information sources. It is proposed to consider the following algorithm of actions necessary for qualitative selection of sources of information based on the analysis of the citation network of scientific publications. A possibility of applying a information sources ranking on a list of "important" articles of a citation network, is reviewed. Existing algorithms of clustering graphs are reviewed and analyzed. A formal description of the network is given and parameters of centrality of citation network nodes are presented. A method of calculating a degree of centrality according to relevance of network nodes, with importance and weight of a series of parameters that are required for publications ranging articles of the citation network. A mathematical model of selecting documents for a library collection on a list of "important" articles of a citation network, is reviewed.

1. Introduction

Research of the state and laws of development of science by means of analysing article citation networks (ACN) is becoming increasingly popular. Searching for information sources to be studied, processed and used in a research work is one of the key stages. Based on the results of an ACN study, one may solve a variety of applied problems, e.g., problems of scientometrics; optimizing the process of building science library collections; preparing science reviews of information sources; compiling literature for research, etc.

The main tool for network specialists is network analysis, which is a set of methods used to visualize networks, describe certain characteristics of the network structure as a whole, as well as obtain detailed information about individual nodes, links and subgroups within the network, create mathematical and statistical models of network structures and network dynamics[1].

Social systems presented in networks are usually characterized by a complex structure. Many social networks consist of relatively dense subgroups (communities), which in turn are connected to each other by weaker ties [2]. Being able to identify and find such subgroups is, therefore, very important. For instance, citation networks may cover different branches of knowledge that contain many articles on specific topics that can be subdivided into subgroups. The simplest type of subgroup is a component.

The most important parameters of the citation network node are its centrality parameters (degree of importance). They characterize the impact of a particular node on the network. When these parameters are obtained, one can the range database articles (DB) according to the importance of the node.

2. Using network analysis for solving practical tasks

Selection of literature by science libraries based on the analysis of citation networks can serve as a vivid example of the practical use of the results of network research. Science libraries are libraries that are designed to facilitate the scientific research process. Taking into account its social designation, a

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

library that satisfies the information needs that have arisen in the process of research activities can be called a science library.

In special literature, the process of building library collections is divided into several stages [3]. Initial selection of publications (documents) is the most time-consuming stage that determines the future effectiveness of the collection. During the initial selection, decisions are made on the acquisition of certain documents for the library collection and on the number of copies to be acquired. At this stage, a document goes through several steps of selection: selecting on the basis of formal grounds, selecting on the basis of semantic correspondence of the document to the collection profile, and selecting on the basis of internal and external correspondence to the system.

A new approach to the selection of documents on formal grounds was proposed in the article "Mathematical model of the optimal choice from a selection of potential documents for the acquisition for a library collection" [4]. The next step of document selection, involving the semantic analysis of the publication content, is the most important part of the selection process, since one must identify the documents according to the criteria of novelty, relevance, originality, fundamentality, reliability, etc. This issue is covered in detail in the article [5].

For the second step, i.e., selecting publications according to semantic correspondence, it will be advisable to apply a selection mechanism based on the results of the research of the article citation network. Studying the citation network in terms of ascertaining the relevance of the articles to certain fields of knowledge will be very useful in finding out which publications are referred to by their authors. The obtained information can be used when acquiring such publications or subscribing to journals in the important industries for the science library collection. This approach will help to solve the problem of acquiring not only new publications issued over the past 3 years, but also publications that are still relevant, but are not included in the publishers' price lists received by acquisition departments of science libraries.

There is a need to optimize the process of building science library collections at the step of selecting documents based on their semantic correspondence to the library collection profile, since it is a challenging task for library acquisition specialists to complete without the assistance of professional experts (specialists competent in the relevant fields of knowledge). It is advisable to optimize this process through the development of a mechanism of publication selection based on the analysis of the results obtained in a study of the article citation network.

It is proposed to consider the following algorithm of actions necessary for qualitative selection of sources of information (literature) based on the analysis of the ACN, which will consist of the following steps:

1) creation of a model of scientific publications citation network;
2) dividing the ACN into thematic sub-clusters (clustering);
3) determining articles that will be " important " to a particular subject;
4) getting a list of articles and ranking it by the level of importance of network nodes;
5) identifying "important" publications referred to by the authors of the " important" articles on the basis of an ACN analysis;
6) ranging the obtained list of publications by the level of importance taking into account data deterioration;
7) selecting publications on the basis of the list of publications bearing in mind the requirements set and the restrictions imposed.

This algorithm is planned to be implemented with the help of software tools that will make up an automated decision support system (ADSS) assisting in making decisions related to building a library collection. The creation of ADSS will help to solve problems that one might come across during the process of document acquisition, e.g., dealing with the complex and time-consuming nature of document selection, avoiding re-ordering the same publications on price lists of new publishers, or re-ordering publications on the same price lists before receiving the initial order. Often, due to the inability to fully assess the semantics of a document, acquisition specialists base their opinion just on the formal features of a document, which can also be avoided through the introduction of ADSS.

The list of tasks to be solved by the proposed ADSS includes:

- prompt presentation of information on publications available in the publishers' catalogues;
- prompt provision of full information on the orders for publications made;
- evaluating and monitoring orders.
- viewing the list of publications in the order of descending priority;
- including in the acquisition list only those publications that meet the specified requirements.

3. Definition of the article citation network

D. Price was the first to define the term "citation network" [6]. He developed a mathematical theory of the growth of these networks and formulated the Price Law on the aging of scientific literature. The works of V.N. Zadorozhnyi and E.B. Yudin also made a significant contribution to the global practice of network research and to the development of methods of network parameter determination [7, 8]. The works of the staff of the Institute of Computational Mathematics and Mathematical Geophysics (Russian Academy of Sciences, Siberian Branch) are also of great practical importance for the study of article citation networks. The experience of these scholars will be used in the study of ACN for the purpose of its application in the library field.

The subject of study is the article citation network. The network is represented as a digraph. Nodes of the network are research publications, and the edges (arcs) are the communications between them, which are carried out through citation. Citation references can be both to research articles (i.e., journal publications indexed in the database) and to other scientific literature, which we will call documents. Then there are a set S containing a certain number of articles (n) and the relation R given by $S \times S$:

$$s_i \, R \, s_j \equiv s_i \text{ cites } s_j \,, \tag{1}$$

determining the citation network.

For articles s_i and s_j in the ratio (1), we accept following provisions:

1) there is no s_i citing itself;
2) if s_j cites s_i, then s_i does not cite s_j.

ACN is represented as a digraph $G = (V, E)$ with the adjacency matrix $M = ||m_{ij}||$. Thus, if $(j, i) \in E$ (article j cites article i), then $m_{ij} = 1$ and $m_{ij} = 0$ otherwise. In graph G the vertices correspond to the articles, and the set E consists of arcs.

4. ACN clustering

Clustering is the first step in the analysis of article citation networks. At this stage, the set of elements is divided into groups (communities). Clustering of ACNs is required in order to attribute particular article segments to a specific branch of knowledge and then determine the "degree of significance" of the node (article) in the cluster. Based on the results of cluster analysis we can reasonably judge about the research directions presented in the analyzed database.

The authors of the article [9] have already noted that the task of clustering such objects for the purposes of bibliometric analysis is substantially complex, since the ACN is a digraph. First of all, this has to do with the fact that most of the developed algorithms are designed for clustering undirected graphs.

One way to solve the problem of clustering the nodes of the ACN into groups is to transform the citation digraph with the adjacency matrix into an undirected graph. Regarding the application of this approach in bibliometry [10], it is proposed to consider the symmetric matrix $M^{coc+bbc} = MM^T + M^TM$, adjacency matrix of the graph, which includes the adjacency matrix $M^{coc} = MM^T$ of the co-citation graph[1] and $M^{bbc} = M^TM$ of the bibliographic combination graph. Thereby an attempt is made to take into account the semantics of articles — both the internal similarities and the external features, such as bibliographic connection and the power of joint citation. After digraphs have been converted to weighted undirected graphs, clustering algorithms designed for the undirected graph are used. The analysis of two different clustering algorithms performed in the *igraph* package is presented in the

[1] The method of co-citation is a method of analysis and study of scientific creativity, which involves identification of references to the works of another author in a research work being analyzed (Dictionary of the History of Psychology, 2007).

work [9]. This paper substantiates the choice of the proposed method by the fact that the issue of clustering is better researched for undirected graphs in terms of implementation algorithms.

Another method of ACN clustering is presented in the work [11]. Based on the clustering algorithm of Business System Planning (BSP), the algorithm of cluster analysis of the directed graph (social network) is proposed. It divides the social network into different classes according to the objects in the social network and the connections between the objects, and can also identify the relations between the clusters. The algorithm is based on determining the reachability of links between objects and contains the following steps:

1) generation of the edge-generating matrix and the pointed matrix;
2) step-by-step calculation of the reachability matrix between objects;
3) calculation of the multi-step reachability matrix between objects;
4) calculation of the reachability matrix;
5) determining the relations between the classes.

However, in this algorithm edges between objects have the same weight (in the actual situation such edges can have different weights), and the properties of each cluster were not analyzed. This indicates the incompleteness of the proposed algorithm of cluster analysis. The main drawback of this algorithm is that it uses matrices to store edges and available relations, whereas in a real social network these matrices will be too large and cannot be loaded into the main memory. The article [12] proposes an improved BSP clustering algorithm for social network analysis. However, there are drawbacks to this algorithm as well. Edges between objects have the same weight, whereas in the actual situation such edges may have different weights.

It should be noted that there are six community allocation algorithms in the igraph library that can be used for ACN clustering:

- algorithm "walktrap.community";
- algorithm "label.propagation.community";
- algorithm "spinglass.community";
- the algorithm of the "leading.eigenvector.community";
- algorithm "fastgreedy.community";
- algorithm "edge.betweenness.community".

The article [13] provides an analysis of the above algorithms. According to the results of the analysis, two algorithms — walktrap.community and fastgreedy.community — are the most relevant for the identification of communities in relation to their applicability in research of large networks. The fastgreedy.community algorithm is designed to analyze large amounts of data, and in this sense it is the most promising one. However, it does not work with directed graphs. Besides, after having allocated the communities, in order for this algorithm to be applied to real networks one needs to solve the problem of identification of vertices, which are indicated here only by numbers. Based on the results of the analysis, one can state that the walktrap.community algorithm reflects the visual representation of the network quite well and gives a list of the vertices attributable to the communities. However, it has a disadvantage: it can be used only if the size of the final network does not exceed several hundred thousand vertices.

5. Centrality parameters of ACN nodes

The next step in the ACN analysis is to identify " important" articles (influential nodes). The most important parameters of the citation network node are its *"centrality"* parameters (degree of *importance*). They characterize the impact of a particular node on the network. When these parameters are obtained, one can the range database articles (DB) according to the *importance* of the node. The concept of "centrality", approaches to the formulation of the basic requirements for the measure of centrality, and its properties are discussed in detail in the article by N. D. Shcherbakova "Axiomatics of Centrality in Complex Networks" [14].

First of all, it is important to determine which parameters of the ACN node characterize its impact on the network, and are sufficient to determine its "centrality". The article [15] defines the following parameters of "centrality" of the nodes of the article citation network:

- degree centrality;
- closeness centrality and harmonic closeness;
- betweenness centrality;
- authority centrality and hub centrality.

Degree centrality. The ACN network is oriented, therefore, incoming and outgoing connections are to be analyzed separately, for which purpose the number of edges that begin or end at this node must be calculated. Accordingly, the *incoming degree* (*in-degree*) is calculated as the sum of the line:

$$indeg(i) = \sum_j a_{ij} \ , \tag{2}$$

and the *outgoing degree* (*out-degree*) is calculated as the sum of the column:

$$outdeg(i) = \sum_j a_{ji} \ . \tag{3}$$

Closeness centrality and harmonic closeness. The first parameter C_C (*closeness*) characterizes the proximity of the node to the rest of the network. As applied to digraphs that are not strongly related:

$$C_c(i) = \frac{1}{\sum_{d_{(i,j)<\infty}, i \neq j} d(i,j)} \ , \tag{4}$$

if vertex *j* is unattainable from vertex *i* , then *d(i, j)* = ∞, so such vertices are excluded from consideration. The *closeness* parameter can be considered both in the *out* mode and in the *in* mode.

For such graphs it is possible to calculate the parameter C_{HC} (*harmonic closeness*), when instead of the inverse of the sum of distances the sum of the inverse distances is taken (assuming that $\infty^{(-1)} = 0$):

$$C_{HC}(i) = \sum_{j \neq i} \frac{1}{d(i,j)} = \sum_{d(i,j)<\infty, j \neq i} \frac{1}{d(i,j)} \ . \tag{5}$$

Betweenness centrality. The parameter C_B (*betweenness*) is the share of shortest paths between all network nodes that pass through this node [15]. The index $C_B(v)$ for vertex *v* is defined as follows:

$$C_B(v) = \sum_{i \neq v \neq j \neq i \in V} \frac{\sigma_{ij}(v)}{\sigma_{ij}} \ , \tag{6}$$

where σ_{ij} is the number of shortest paths from vertex *i* to vertex *j* of the graph, and $\sigma_{ij}(v)$ is the number of shortest paths from *i* to *j*, passing through *v*.

Authority centrality and hub centrality. The article is authority if it has authoritative content on a particular topic, which is confirmed by repeated citation of the article. Nodes formed by such articles (*AU*-articles) are called *authorities*. The article is an informative article (*hubs* (*HU*-article)) if the list of references contains many references to *AU*-articles. Nodes that are formed by *HU*-articles are called information nodes.

The value of the parameter C_{AU} (influence of the node *i*) is calculated as follows:

$$C_{AU}(i) = \alpha \sum_j a_{ij} C_{HU}(j) \ , \tag{7}$$

The values of the parameter C_{HU} (information value of the node *i*) are calculated as follows:

$$C_{HU}(i) = \beta \sum_{j} a_{ij} C_{AU}(j) \ , \tag{8}$$

where α and β are constants. A description of the method for calculating the values of these parameters is given in the work [15].

A summary measure of centrality by importance can be measured through a complex assessment of the centrality parameters of the ACN nodes, represented as the sum of the ranks of the network nodes for each indicator of centrality, taking into account their significance and weight. The value of the parameter C_{sum} (*important*) is calculated as follows:

$$\mathrm{C}_{sum}(i) = \sum_{i=1}^{n} k_i R_i \ , \tag{9}$$

where R_i is the rank of the *i*th indicator;

n is the number of parameters;

k_i is the coefficient of weight of the *i*th indicator.

6. Selection of information sources based on the list of " important" articles

To implement the algorithm of selecting information sources proposed in paragraph 1, the principle of ranging the resulting list of references by the number of references to publications based on the study of the ACN is introduced.

It is proposed to consider the following mathematical model of publication selection. For each publication D_i from the list of sources of "important" articles of one community (cluster) of the database on the subject *t*, the local citation index $I\ (D_i, T)$ is calculated as the total number of references to this source of the same field of knowledge. Further, any publication having the name D_i is ranged by significance for this thematic subgroup by summing up such indices for all the sources from the list of the literature used in the articles. This principle of ranking by citation index is proposed in [16].

In accordance with this principle, we introduce a popularity index of a publication is the sum of its local citation indices:

$$I_{sum}(i) = \sum_{j=1}^{n} I\left(D_i, T_j\right), \tag{10}$$

where n – number of topics T_j ,

T_j – topics of publications referring to the publication D_i .

The overall rating calculated for each publication from the list of "important" articles allows us to assess the degree to which scholars (library users) require the selected publications and thus to make decisions on orders of publications for the library collection.

7. Conclusion

Based on the results of the analysis of the network, we can reasonably judge about the research directions presented in the articles. The proposed method of measuring the centrality of network nodes (9) enables one to conduct a single-value ranging of ACN articles.

The complex nature of building science library collections, as well as the fact that the acquisition specialists have to manually create an order for publications to be included in the library collection, based on the results of their subjective opinion, leads to a situation where the order to be obtained is not the optimal choice. Therefore, one can conclude that it is necessary to develop a mechanism that would support decision-making regarding the selection of publications for the library collection, based on objective assessments. To support this decision, optimization of the process publications selection

through the analysis of article citation networks is proposed, and the corresponding mathematical models presented in this article are constructed.

It should be noted that this algorithm for the selection of sources of information based on the analysis of the ACN, presented in the article, can be applied in practice in solving any problems related to the search for quality sources of information.

As of today, there are no automation tools that would provide support for decision-making regarding the choice of publications and making an optimum order, taking into account the current restrictions. Therefore, the development of ADSS based on the mathematical models proposed in this article is a topical and necessary task.

The problem of choosing a clustering algorithm remains open, as the available algorithms to some extent have their drawbacks in relation to directed graphs.

The proposed approach to the formation of lists of "important" publications can be applied in the following:

- creating lists of reference for theses;
- selecting literature for research;
- preparing scientific literature reviews;
- in the development of training manuals and guidelines;
- solving scientometric problems;
- etc.

Considering the large number of essential problems in the field of science that need to be addressed, one may highlight the relevance of and the need for the research of article citation networks.

References

[1] Luke D A 2017 *A User's Guide to Network Analysis in R* (Moscow: DMK Press) p 250
[2] Granovetter M 2009 The Strength of Weak Ties' *Economic Sociology* (*Electronic Materials* vol 10 no 4) pp 31-50
[3] Stolyarov Y N 2015 *Library Collection* (Saint Petersburg: Professiya) p 384
[4] Olgina I G 2018 Mathematical model of optimal choice from a set of potential documents for building a library collection *Mathematical and Computer Modeling: collection of materials of the 6th International Science Conference* (Omsk: OmSU) pp 113-115
[5] Vikhreva G M 2006 Technology of assessing documents for the purposes of their adding to collections of large science libraries *Library collections: Problems and Solutions (Electronic Materials* no 9)
[6] Price D 1965 Networks of scientific papers *Science* vol 149 no 3683 pp 510-515
[7] Zadorozhnyi V N, Yudin E B 2019 *Vertex degree distribution and arc endpoints degree distribution of graphs with a linear rule of preferential attachment and Pennock graphs* https://arxiv.org/abs/1904.00426
[8] Yudin E B 2012 Methods of structural identification of stochastic networks and random graph generation in the problems of complex systems modeling *Thesis of a Candidate of Technical Sciences* (Omsk) p 150
[9] Satuluri V, Parthasarathy S 2011 Symmetrizations for clustering directed graphs Proc. 14th Internat. Conference on extending database technology (*Electronic Materials*) (Uppsala, Sweden) pp 343-354
[10] Bredikhin S V, Lyapunov V M, Shcherbakova N G 2016 Structure of the citation network of scientific articles *Journal of Computer Science Issues* no 3 pp 26-43
[11] Yu G Social 2007 Network Analysis Based on BSP Clustering Algorithm *Communications of the IIMA* vol 7 no 4 pp 39-45
[12] Sanjiv Sharma, Gupta R K 2010 Improved BSP Clustering Algorithm for Social Network Analysis *International Journal of Grid and Distributed Computing* vol 3 no 3 pp 67-76
[13] Kincharova A 2012 *Application of community detection algorithms for sociological research of blogs: results of the pilot study* (*Electronic Materials*)

[14] Sherbakova N D 2015Axiomatics of centrality in complex networks *Problems of Computer Science* no 3 pp 3-14
[15] Bredikhin S V, Lyapunov N G, Shcherbakova N G, Yurgenson A N 2016 Parameters of the "centrality" of nodes in an article citation network *Problems of Computer Science* no 1 pp 39-57
[16] Kromina L A 2012 Automated decision support based on the ranging of publications according to the level of need for a university library ordering literature *Thesis of a Candidate of Technical Sciences* (Ufa) pp 200