Predicting the dynamics of scientific activities: A diffusion-based network analytic methodology

Conference Paper · November 2018

CITATIONS
0

READS
29

4 authors, including:

Yi Zhang
University of Technology Sydney
53 PUBLICATIONS **409** CITATIONS
SEE PROFILE

Ximeng Wang
Beijing Jiaotong University
9 PUBLICATIONS **29** CITATIONS
SEE PROFILE

Please cite as:

Zhang, Y., Wang, X., Zhang, G., Lu, J. 2018. Predicting the dynamics of scientific activities: A diffusion-based network analytic methodology, *the 81th Annual Meeting of the Association for Information Science and Technology*, Vancouver, Canada.

Predicting the dynamics of scientific activities: A diffusion-based network analytic methodology

Yi Zhang
Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia. Yi.Zhang@uts.edu.au

Guangquan Zhang
Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia. Guangquan.Zhang@uts.edu.au

Ximeng Wang
Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia. Ximeng.Wang@student.uts.edu.au

Jie Lu
Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia. Jie.Lu@uts.edu.au

ABSTRACT

With the rapid explosion of information and the dramatic development of bibliometric techniques in the past decades, it becomes a challenge to comprehensively, extensively, and efficiently understand science maps. Aiming to explore in-depth insights from science maps and predict the dynamics of scientific activities, this paper, based on the co-occurrence statistics of terms derived from scientific documents, proposes a diffusion-based network analytic methodology to conduct the prediction study from two aspects: the research interest of scientific researchers and the evolutionary directions of scientific topics. A case study on academic articles downloaded from three leading journals in the field of bibliometrics demonstrates the feasibility of the methodology. The future directions of bibliometrics are identified, such as the application of information technologies to traditional bibliometric data, the interactions between bibliometrics and science, technology, and innovation policy issues, and individual-level bibliometrics. The results also provide recommendations as potential research interests for a set of experts. The proposed method could be a toolkit to conduct forecasting studies for a given technological area or a given discipline, and a recommender system to assist academic researchers in identifying potential research interests and extended areas.

KEYWORDS

Bibliometrics, network analysis, diffusion, scientific activities.

INTRODUCTION

Science maps are described as the graphic reference systems representing the relationships among scientific disciplines and portfolios (Small 1999; Börner 2014), and have become one crucial part of bibliometric studies these years. Related endeavours include: introducing multiple bibliometric indicators (e.g., citation statistics, words/terms, bibliographic coupling, and specific classification codes) or their combinations for accurate knowledge representation (Noyons et al. 1999; Glänzel 2001; Chen et al. 2010; Klavans & Boyack 2017), applying information technologies (e.g., text analytics, topic models, and machine learning) to develop advanced bibliometric models for multiple purposes (van Eck et al. 2010; Ding & Chen 2014; Zhang et al. 2017), and integrating science maps with a broad range of tools (e.g., Google Map), scientific databases, and actual cases (Leydesdorff & Bornmann 2012; Huang et al. 2016). Unfortunately, with the rapid explosion of information, the structure of science maps are becoming more and more integrative and complicated, and thus comprehensively and extensively understanding science maps challenges both bibliometric communities and related users. Under this circumstance, one concern is how to explore the interactions between elements on science maps and predict underlying relationships that may appear in future.

Benefiting from the development of information technologies, network analysis provides opportunities to address the concern. Complex network analysis was first raised by physics communities, in which complex systems in nature and society are described as networks and many methods were proposed to reveal network structures and the relationships among their units (Palla et al. 2005). Complex networks are known as social networks in bibliometrics, especially when investigating the relationships between authors, portfolios (e.g., academic institutions and commercial firms), or journals (Otte & Rousseau 2002), and the comparison between complex networks and science maps have also been fully examined (Hung & Wang 2010; Yan et al. 2010). As a mainstream network analytic approach, centrality measures, first developed by Freeman (1977) for evaluating the structure of social networks, have been well-upgraded and applied to analyze co-authorship-based collaboration networks

(Yan & Ding 2009; Abbasi et al. 2012). Despite the fact that these efforts provide new angles to further understand science maps, the way to predict the dynamics of scientific activities by profiling historical data is still elusive.

Aiming to address these considerations, this paper, based on the co-occurrence statistics of terms derived from scientific documents, is to propose a diffusion-based network analytic methodology to predict the dynamics of scientific activities from two aspects: the research interest of scientific researchers and the evolutionary directions of scientific topics. On one hand, considering unsupervised environments in a large number of real-world cases, we introduce an indicator *divergence* to a K-means-based clustering approach to automatically decide the number of topics in a given interval. On the other hand, with the aid of science maps, bipartite networks are used for constructing topic-term networks and author-term networks, and a diffusion-based analytic approach is proposed to predict the dynamics of scientific activities by detecting the diffusion trend of resources in the networks. A case study with 7,258 academic articles downloaded from *Journal of the Association for Information Science and Technology* (JASIST[1]), *Journal of Informetrics* (JOI), and *Scientometrics* (SCIM) was conducted to demonstrate the feasibility of our method, and the results hold interest to not only bibliometric communities but also a wide range of audiences in information science and technology, science policy, and industrial sectors.

The structure of this paper is organized as follows: the following section proposes the diffusion-based network analytic methodology, the application of our methodology to a set of bibliometrics-related articles follows, and finally we conclude limitations and future study.

METHODOLOGY

The framework of the diffusion-based network analytic methodology is given in Figure 1, including two models: term-based topic identification and diffusion-based network analysis.

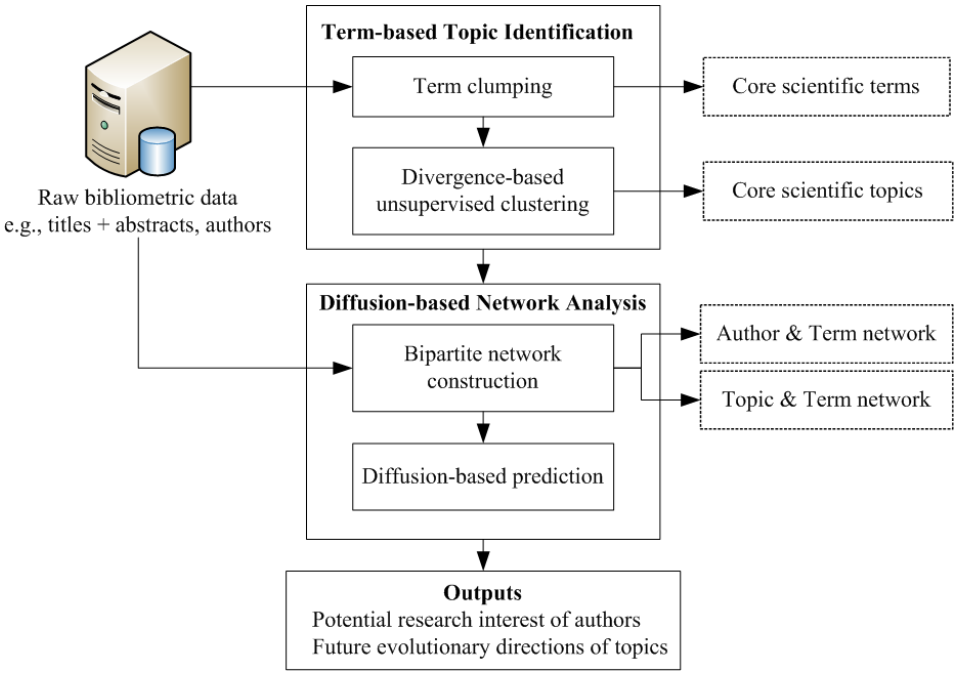

Figure 1. Framework of the diffusion-based network analytic methodology.

Input

The input of our methodology is raw bibliometric data derived from academic databases, e.g., Web of Science (WoS), and mainly includes titles, abstracts, and author information. With the use of VantagePoint (VP)[2], we will combine titles and abstracts and use the natural language processing (NLP) function of VP to retrieve terms. At the same time, some author name disambiguation (AND) techniques incorporated in VP can be applied to clean author information.

[1] We also involve previous articles published by JASIST, with its old name *Journal of the American Society for Information Science and Technology*.
[2] VantagePoint is the software for bibliometric data-oriented text mining and visualization. More details can be addressed on the website: https://www.thevantagepoint.com/

Term-based Topic Identification

This step is to identify core scientific terms and topics, which includes a term clumping function and a divergence-based unsupervised clustering function. Despite not the main focus, we will also pre-process author information in this model.

Aiming to remove noise and consolidate synonyms, the term clumping function integrates stopword-based thesauri and knowledge-based rules in a semi-automated stepwise process (Zhang et al. 2014). It holds the ability to reduce the scale of terms and identify the remaining thousand terms as core scientific terms.

Considering most real-world bibliometric data is unlabeled and thus it is difficult to construct a training set for training parameters (e.g., the number of topics), we specifically integrate an indicator *divergence* with a K-means-based clustering algorithm (Zhang et al. 2016). This indicator can help automatically decide the best number of topics in a given interval, adapting to unsupervised environments. The indicator *divergence* is defined to measure the difference between topics, i.e., a higher divergence means larger differences exist between two topics, and the performance of such clustering can be better. We follow the idea that introducing the Jensen-Shannon divergence (JSD) approach (Lin 1991) to measure the distance between the word probability vectors of two documents (Boyack et al. 2011), but we calculate the distance between the term frequency vectors of two topics.

We denote $T(\Phi)$ as a topic and $C(\Psi)$ as the entire data corpus, of which Φ and Ψ are the term vectors. The frequencies of a term θ in topic $T(\Phi)$ and corpus $C(\Psi)$ can be described as θ_T and θ_C respectively. Given that α and β are the proportions of a term θ in topic $T(\Phi)$ and corpus $C(\Psi)$ respectively, i.e., $\alpha = \theta_T/||\Phi||_1$ and $\beta = \theta_C/||\Psi||_1$.

$$||\Phi||_1 = \sum_{i=1}^{dim(\Phi)} |\theta_i| \tag{1}$$

$$||\Psi||_1 = \sum_{i=1}^{dim(\Psi)} |\theta_i| \tag{2}$$

where $dim(\Phi)$ and $dim(\Psi)$ are the dimensions of vectors Φ and Ψ respectively.

Let$v = (\alpha + \beta)/2$, and the Kullback-Leibler divergence$D_{KL}(\alpha, v)$ and $D_{KL}(\beta, v)$ can be calculated as:

$$D_{KL}(\alpha, v) = \sum_{i=1}^{dim(\Phi)} \alpha_i \times log(\alpha_i/v_i) \tag{3}$$

$$D_{KL}(\beta, v) = \sum_{i=1}^{dim(\Psi)} \beta_i \times log(\beta_i/v_i) \tag{4}$$

Then, the JSD value of topic $T(\Phi)$ is:

$$JSD(T(\Phi)) = \frac{1}{dim(\Phi)} \frac{D_{KL}(\alpha,v)+D_{KL}(\beta,v)}{2} \tag{5}$$

Compared with the JSD formula used by Boyack et al. (2011), we take the dimension of the vector Φ into consideration to eliminate possible negative influence raised by the size of topics, e.g., a topic with more terms will have a higher JSD value. Furthermore, since the JSD value is calculated for each topic, we set the average JSD value of all topics as the divergence value of the current-round cluster analysis. Generally, when deciding the number of topics in a given interval, we prefer to choose the one with the highest divergence value. Thus, topics generated by the divergence-based clustering function are identified core scientific topics, and together with core scientific terms, are considered as the outputs of this step.

Diffusion-based Network Analysis

This step is to predict the dynamics of scientific activities, i.e., the potential research interest of researchers and the future evolutionary directions of topics, and it includes a bipartite network construction function and a diffusion-based prediction function.

As a particular network, the nodes of a bipartite network consist of two sets and edges only exist between two nodes in different sets (Zhou et al. 2007). Academic researchers (one set of nodes) that focus on similar research interests can be connected by terms (the other set of nodes), and the same as scientific topics and terms. Under this circumstance, it is reasonable to consider author-term networks and topic-term networks as bipartite networks. Using the author-term network as an example, we describe the algorithm of bipartite network construction and diffusion-based prediction as follows:

We denote an author-term network as $G_{A\&T}(A, T, L)$, where $A = \{a_1, a_2, \cdots, a_m\}$ is a set of authors, $T = \{t_1, t_2, \cdots, t_n\}$ is a set of terms, and $L = \{l_1, l_2, \cdots, l_z\}$ is a set of edges that link authors and terms. Generally, the bipartite network $G(A, T, L)$ can be represented as an $m \times n$ adjacency matrix $C_{A\&T}$, of which a cell $c_{i\alpha}(1 < i < m; 1 < \alpha < n)$ indicates the value of the edge between node t_α and node a_i, and

$$c_{i\alpha} = \begin{cases} 1 \text{ if term } \alpha \text{ appears in at least one article of author } i \\ 0 \text{ if term } \alpha \text{ does not appear in any article of author } i \end{cases} \quad (6)$$

An assumption in bipartite network projection is given that a number of resources are associated with one set of nodes in the network, and these resources can diffuse between two different sets of the bipartite network (Zhou et al. 2007). Thus, using a diffusion strategy that resources will be equally distributed by the degree of linked nodes (i.e., the number of edges connecting with the node), a mass diffusion-based analytic technique (Wang et al. 2016) was designed to detect such diffusion. Given that we plan to predict the potential research directions of author a_i, i.e., new terms that author a_i will collect in the near future. The diffusion value $f(a_i, t_\alpha)$ is defined as the final resource that term t_α will obtain from other terms for author a_i, and can be calculated as:

$$f(a_i, t_\alpha) = \sum_{j=1}^{m} \left(\frac{c_{j\alpha}}{k(a_j)} \sum_{\beta=1}^{n} \frac{c_{i\beta} c_{j\beta} * f(t_\beta)}{k(t_\beta)} \right) \quad (7)$$

where $c_{j\alpha}$, $c_{j\beta}$ and $c_{i\beta}$ are the elements in adjacency matrix $C_{A\&T}$, $k(a_j)$ and $k(t_\beta)$ are the degrees of author j and term β respectively, and $f(t_\beta)$ is the initial resource on term β.

An example of such diffusion in a bipartite network is given in Figure 2, where nodes a_1, a_2 and a_3 represent authors, and nodes t_1, t_2, t_3 and t_4 are terms. Given that we target to author a_1, and terms t_1, t_2 and t_3 appear in the articles published by author a_1, thus, the three terms get 1 unit of resources respectively. In Step 2, these resources are equally distributed to authors by the degrees of the terms (e.g., author a_1 receives 1/2, 1, and 1/3 units of resources from terms t_1, t_2 and t_3 respectively). Then, authors flow their obtained resources back to their linked terms, and the diffusion value of each term for author a_1 is the final resource that it collects in Step 3, e.g., the diffusion value of term t_1 for author a_1 is 8/9.

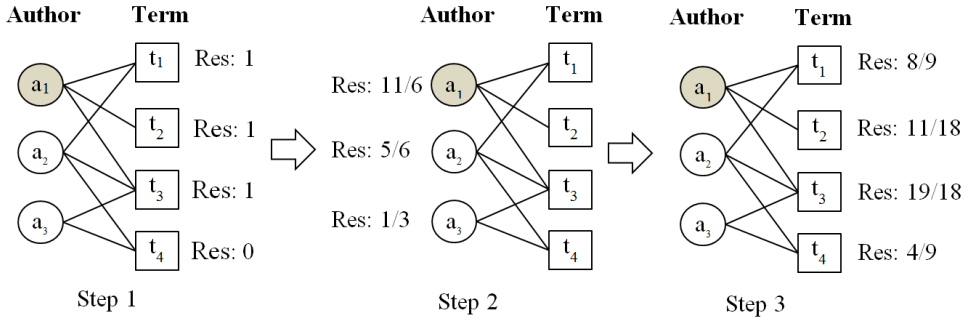

Figure 2. Sample of diffusion between nodes in a bipartite network.

Based on Equation 7, author a_i will have a diffusion vector $v(a_i)$ indicating the final resources that the author can obtain from all terms, and all these vectors will constitute a diffusion matrix $D_{A\&T}$, which can be considered as one output of our methodology. Furthermore, regarding the task of predicting the potential research interest of an author, we will remove all terms that have ever appeared in the articles of author a_i, i.e., $c_{i\alpha} = 1$, and rank the remaining terms by the diffusion value. At this stage, those terms that have never appeared in an author's articles but are ranked in the top of the list can be the potential research interest of the author. Following the same steps, we can also identify the terms that indicate the future evolutionary directions of topics. Therefore, a list of terms that hold the capability to predict the dynamics of scientific activities will be the final output of our methodology.

CASE STUDY

Aiming to demonstrate the feasibility and efficiency of our methodology, we conducted a case study to articles published by three leading journals in the areas of bibliometrics, i.e., JASIST, JOI, and SCIM, and the outputs include: the potential research interest of our bibliometric researchers and the future evolutionary directions of bibliometrics-related topics. The search was on WoS database, applied on March 22, 2017, and the collected dataset contains 7,258 articles and covers the time period from 2000 to 2017[3]. The detailed information (including the time coverage and the number of articles) of the three journals is given in Table 1.

Journal	Time Coverage	Number of Articles
JASIST	2000-2017	3225
JOI	2007-2017	698

[3] The time coverage of the *Journal of Informetrics* is from 2007 (its established year) to 2017.

SCIM	2000-2017	3335

Table 1. Detailed information of the three journals.

Term-based Topic Identification

113,159 terms were retrieved by an NLP function in VantagePoint[4], and the term clumping function was applied to identify core scientific terms. The stepwise results are given in Table 2, and the 9,942 terms generated in Step 8 is considered as core bibliometrics-related terms.

Step	Description	# Terms[1]
0	Raw terms retrieved by the NLP technique;	113,159
1	Remove terms starting/ending with non-alphabetic characters, e.g., "1.5%";	98,943
2	Remove meaningless terms, e.g., pronouns, prepositions, and conjunctions;	94,282
3	Remove common terms in scientific articles, e.g., "introduction";	92,942
4	Consolidate terms with the same stem, e.g., singular and plural, and parts of speech;	79,380
5	Consolidate synonyms based on expert knowledge, e.g., "co-word analysis" and "word co-occurrence analysis";[2]	58,624
6	Consolidate terms related to the same country or region, e.g., "Chinese economy" and "Chinese science policy";[3]	57,770
7	Remove terms appearing in only one article;	11,211
8	Remove single words;[4]	9,942

Table 2. Stepwise results of the term clumping.

Note. 1) #Terms – Number of Terms. 2) As bibliometric researchers, we quickly reviewed the results of Step 4 and summarized a list of criteria for consolidating bibliometric synonyms. 3) Considering those country/region-oriented cases have limited relationships with bibliometric methods, we decided to combine such terms to the name of related country/region. 4) According to the study conducted by Zhang et al. (2014), the meaning of single words can be fully covered by related terms, e.g., "fuzzy" and "fuzzy sets". However, as an example, we consolidated "classification analysis" to "classification" in Step 5, but we did not remove single words such as "classification" in Step 8. In addition, several extremely high-frequency terms were also removed, e.g., "bibliometrics", "scientometrics", and "informetrics".

A term co-occurrence map generated by VOSViewer (van Eck & Waltman 2009) is given in Figure 3. As an overall view, certain hotspots are observed: 1) as a mainstream in information sciences, information retrieval occupied a crucial proportion in the three leading bibliometric journals. 2) Led by citation analysis, the development of bibliometric methods is another main focus. 3) Case studies that apply quantitative methods (e.g., bibliometrics and information retrieval techniques) to handle real-world problems in specific countries or regions are another hotspot, and the United States (US) is widely involved.

[4] https://www.thevantagepoint.com/

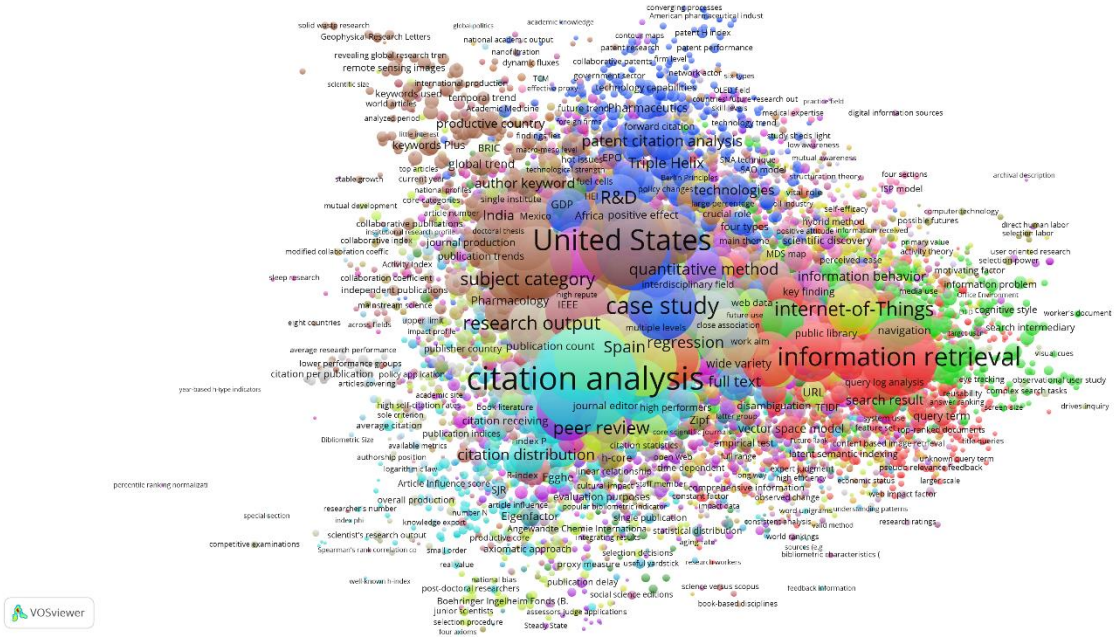

Figure 3. Co-occurrence map for core bibliometrics-related terms (from 2000 to 2017).

An article-term matrix linking the 7,258 articles and the 9,945 terms was constructed, and we ran the divergence-based clustering function by considering the number of topics in the interval [5, 20]. The divergence values of the 16-round experiments are given in Figure 4.

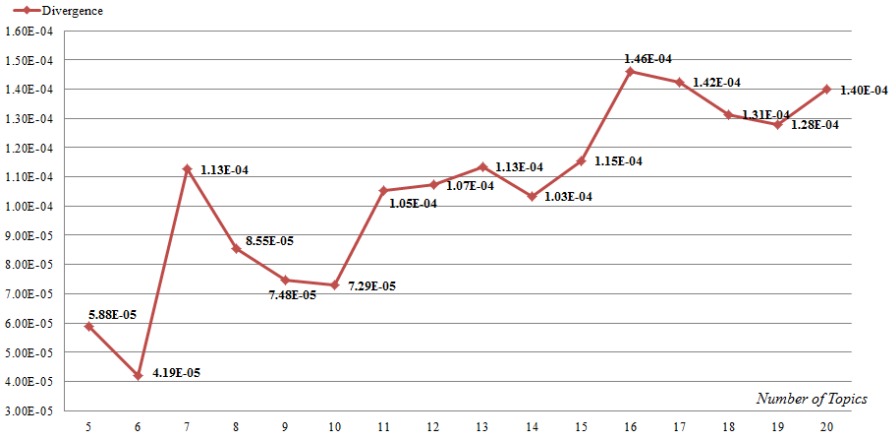

Figure 4. Divergence values of clustering with different number of topics.

Since our criterion of deciding the number of topics is to select the one with the highest divergence value, 16 topics were generated and considered as core bibliometrics-related topics. The detailed information of the 16 topics is given in Table 3.

No.	Label	Description
1	information retrieval	ranking method, search engine, machine learning, digital library, metadata, ontologies, natural language processing;
2	European Union	United Kingdom, R&D, internal collaboration, Spain, EPO, national research system, Italy, Germany, Czech Republic;
3	research evaluation	journal citation report, impact factor, h index, g index, measurement, individual scientist, journal quality;
4	science map	co-word analysis, visualizations, large scale data visualization, cluster analysis, co-citation network, multidimensional scale, robustness, bibliographic data;
5	social network	bibliographic coupling, Pearson correlation, betweenness, centrality measures, correlation analysis, complex network analysis;
6	collaboration research	co-authorship analysis, scientific community, bibliometric indices, research institute, collaboration network;

7	United States	United States, R&D, research institute, North America, technologies, management, public research, population, publication activity;
8	social science	governance, Triple Helix, natural science, patent data, university government industry relations, science & technology, developed country, life science;
9	citation analysis	citation network, co-citation analysis, self-citation, citation rate, citation distribution, altmetric, citation pattern, citation behavior;
10	decision making	research policy, optimization, library management, public policy, clinical medicine, stimulated recall interviews, policy maker;
11	information system	information behavior, computer science, information technology, human computer interaction, user interaction, semantic analysis, information interaction, ontologies, linear model;
12	statistical analysis	correlation analysis, regression, sample size dependency, Poisson distribution, data analysis, computer science, software system;
13	scientific output	bibliometric indices, research group, socioeconomics, female researcher, male researchers, academic collaboration;
14	data mining	data mining, text mining, classification, probability, longitudinal analysis, simulations, prediction model, mathematics, fuzzy set;
15	interdisciplinary research	subject category, interdisciplinarity, applied research, multidisciplinary approach, disciplinary diversity, longitudinal analysis;
16	China	Hongkong, India, collaboration pattern, Tsinghua University, policy implication, emerging economies, comparable analysis, innovative technology, Taiwan;

Table 3. Core bibliometrics-related topics.

As shown in Table 3, the 16 topics and related terms further extend our observations derived from Figure 2 and in particular enrich the details of the three hotspots: 1) the main areas of highly involved information technologies include not only information retrieval but also statistics, data mining, information system, and network analysis; 2) Science map, citation analysis, co-citation analysis, co-word analysis, and co-authorship analysis are highlighted for bibliometrics. 3) Specific real-world issues are also addressed, e.g., research evaluation, collaboration research, interdisciplinary research, and a broad range of science, technology, innovation, and policy (STIP) issues. 4) Three hottest countries/regions are raised, i.e., the US, China, and European Union (EU), and their related countries and regions are also specified. In addition, diverse foci among the three hottest countries/regions are raised, e.g., national research system in EU and emerging economies in China.

Regarding author information, we retrieved 9,555 author names from the 7,258 articles, including both authors and co-authors. A light AND function and a co-authorship-based macro for consolidating low-frequency authors to related high-frequency authors in VantagePoint were used for pre-processing, and 8,692 authors were finalized.

Diffusion-based Network Analysis

Two matrices were constructed for bipartite networks, i.e., a 16×9942 matrix for the topic-term network and an 8692×9942 matrix for the author-term network. The diffusion-based prediction function was first applied to the topic-term network, and based on the diffusion value the prediction for the future evolutionary directions of the sixteen bibliometrics-related topics is given in Table 4.

Topic	Future Research Direction[2]
1	subject category, research evaluation , collaboration research, active research, publication output, *journal citation report*, *R&D*, *policy making*;
2	prediction model, subject category, h index, journal impact factor, uncertainty, *centrality measures*;
3	collaboration network, social science, current study, qualitative method, empirical study, *research trend*, *public health*;
4	population, peer review, uncertainty, empirical result, publication output, *document type*, *impact factor ranking*;
5	computer science, collaboration network, semantic analysis, South Korea, reputation, *comparable analysis*;
6	information seeking, Netherland, subject area, academic journal, web search, *age distribution*, *bibliometric evaluation*, *overlap map*;
7	co-authorship network, research evaluation, machine learning, domain knowledge, bibliometric measure, *dissemination*;
8	full text, scientific community, Google, public health, dataset, *betweenness*, *digital library*;
9	information seeking, multidimensional scale, Lotka's law, knowledge management, domain knowledge, *technologies*, *metadata*;

10	Google Scholar, co-citation network, complex network analysis, research product, semantic analysis, *cluster analysis*, *text mining*;
11	research group, h index, journal impact factor, individual scientist, scientific field, *bibliometric indices*, *reputation*, *scientific output*;
12	longitudinal analysis, empirical result, centrality measure, research area, healthcare, *research policy*, *individual level*, *co-word analysis*;
13	developed country, internet of things, cluster analysis, text mining, wide range, *co-authorship analysis*;
14	growth rate, scientific literature, scientific output, academic research, co-word analysis, *internal collaboration*, *webometric analysis*, *knowledge management*;
15	future research direction, mathematics, research evaluation, research product, bibliometric indices, *individual level*;
16	multidisciplinary analysis, quantitative method, self-citation, individual scientist, scientific field, *United Kingdom*, *content analysis*;
Validation Measure[3]: There are 44 relevant terms, 14 "not so" relevant terms, and 22 irrelevant terms out of 80 terms, so we consider the precision as 0.64 (i.e., $(44 \times 1 + 14 \times 0.5)/80$).	

Table 4. Prediction for the future evolutionary directions of the bibliometrics-related topics[1].

Note. 1) We follow the sequence of the 16 topics given in Table 3. 2) The first five terms are the Top 5 terms with the highest diffusion value to related topics, and based on expert knowledge the italicized terms were selected from the list of the Top 6-10 terms and treated as supplementary terms. 3) We manually evaluated the first five terms of each topic with the following criteria: if a term is irrelevant with bibliometrics, we mark it with a solid underline (as 0 point); if a term is not so relevant, we mark it with a broken underline (as 0.5 point); if a term is relevant, we do nothing but as 1 point.

Observations on Table 4 are summarized: 1) one common interest is to apply information technologies (e.g., information retrieval, data mining, and information systems) to assist in specific bibliometric issues, e.g., evaluating research performance from different levels (e.g., h index for scientists and impact factor for journals), profiling scientific activities for analyzing internal or external collaboration, and decision support for R&D planning and policy making. 2) Individual-level bibliometrics seem to be an emergent direction, which include not only the development of related statistical models but also individual scientist-oriented bibliometric studies. 3) The engagement of text mining techniques with a wide range of mainstream bibliometric approaches (e.g., citation/co-citation analysis, science map, and social network analysis) is highlighted, since terms such as "semantic analysis", "metadata", "full text", and "content analysis" can be addressed here and there. 4) Intriguing terms are raised in specific country or region-based case studies, e.g., uncertainty issues in Europe, knowledge dissemination in the US, and self-citation issues in China.

Before moving our focus to the author-term network, we generated an author correlation network in Figure 5 to briefly spotlight certain leading experts and author groups, since bipartite networks are relatively abstract and not well-designed for visualization (e.g., two sets of nodes are located in columns, as shown in Figure 2). Note that Figure 5 is not a bipartite network, since it contains only one set rather than two sets of nodes, and the correlation coefficient between two authors is calculated by the Pearson measure, based on their 9,942-dimension term vectors.

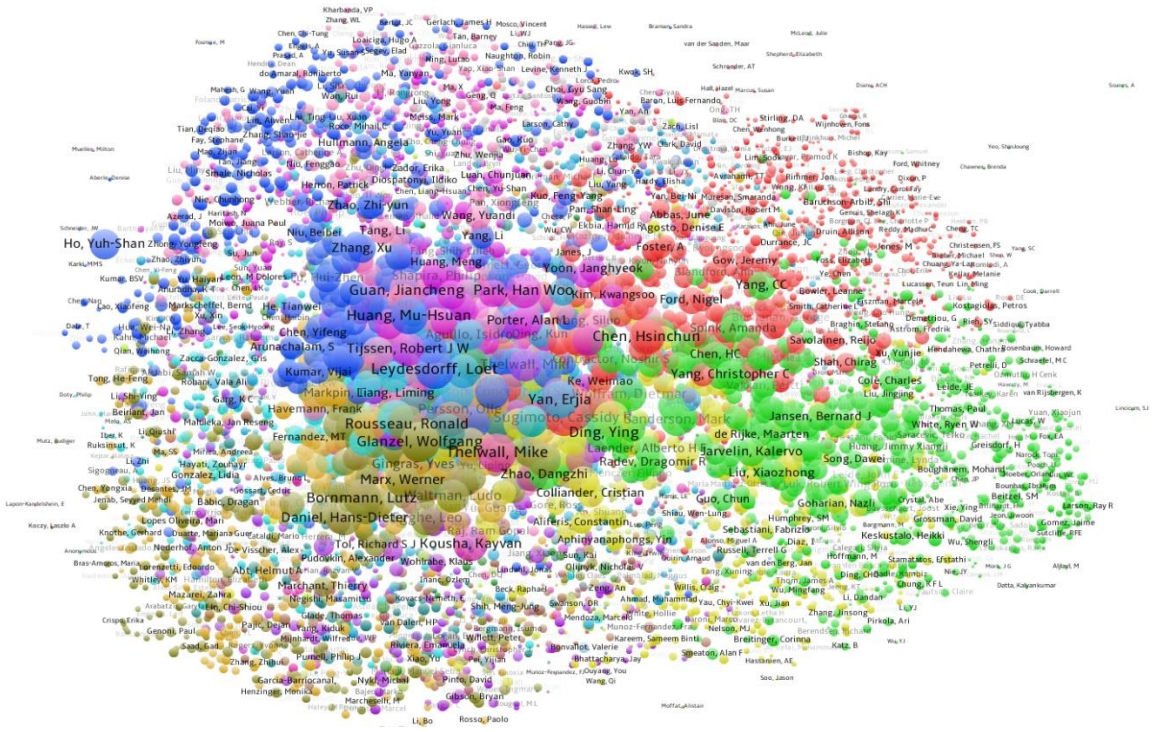

Figure 5. Correlation network of bibliometrics-related authors.

Note. Despite the use of the AND algorithm for data pre-processing, certain duplicate names might still exist in Figure 5, e.g., Chen, Hsinchun and Chen, HC. Since AND is not the main focus of our study, we do not provide any improvements in the current paper but it is definitely an important direction of our future study.

The prediction study on the potential research interest of authors was conducted from two ways: 1) focusing on the potential interest of some leading experts (those who have published a large number of articles covering a wide range of bibliometric topics and have established close interactions with worldwide researchers) to foresee the research frontiers of bibliometric topics, and 2) to select some relatively "niche" authors (those who do not publish so many articles and usually concentrate on a relatively small research circle) to delve into specific research directions. At this stage, five authors were selected, i.e., Loet Leydesdorff, Wolfgang Glanzel, Ludo Waltman, Hanwoo Park, and Alan Porter, and we then applied the diffusion-based prediction function to the author-term network, and the results are given in Table 5.

Author	Potential Research Interest[1]
Leydesdorff, L.	computer science, collaboration network, scientific products, interdisciplinary research, patent citation analysis;
Glanzel, W.	co-word analysis, social network, management, computer science, international collaboration;
Waltman, L.	United States, social science, information retrieval, China, scientific community;
Park, H.	ranking method, text mining, management, scientific community, information retrieval;
Porter, A.	citation analysis, information retrieval, h index, social network, scientific community;

Table 5. Prediction for potential research interests of selected authors.

Note. 1) We choose the terms from Top 10 terms with the highest diffusion value.

As leading experts in the field of bibliometrics, Dr Leydesdorff, Dr Glanzel, and Dr Waltman are ranked as the 1st, 4th, and 12th authors with the largest number of articles published in JASIST, JOI, and SCIM from 2000 to 2017. Based on the articles of these three leading bibliometric experts, we summarize our findings as follows: 1) indicated from those over 150 articles in our dataset, the main foci of Dr Leydesdorff cover a very broad range of bibliometrics, e.g., Triple Helix, social network analysis, science maps, and bibliometric methodology. Considering existing topics and the interest of those who share close similarity with Dr Leydesdorff, we specifically pick up two possible directions: introducing techniques in the field of computer science to enhance the capability of bibliometrics, and applying bibliometric approaches for interdisciplinary or multidisciplinary research. 2) Over 100 articles published by Dr Glanzel are involved, including a large number of topics on information science and library science, e.g., bibliometric indicators, probabilistic models, h index, and interdisciplinary/multidisciplinary studies. Our recommendation includes: the synergistic effect of co-word-based bibliometric approaches and other indicators (e.g., citation statistics), and bibliometrics-based case studies on investigating international collaboration, social network, and management. 3) Compared with those specified approaches or methods recommended to Dr Leydesdorff and Dr Glanzel, it is intriguing

to investigate the research interest of Dr Waltman, whose research mainly focuses on the development of bibliometric indicators, analytic models, and networks. Our methodology considered that Dr Waltman might hold interest to apply his research to address empirical insights on real-world STIP issues, e.g., specific country/region-oriented case studies. As a comparison, we selected two more authors from our list, Dr Hanwoo Park who is a co-author of Dr Leydesdorff on Triple Helix-based case studies and also dedicates to webometrics and social network analysis, and Dr Alan Porter whose research aligns with both public policy and bibliometrics. Similarly, we recommend techniques such as ranking method, text mining, and information retrieval to Dr Park, which can help extend the adaptability and analytic capability of related webometric approaches, when we provide alternative bibliometric approaches (e.g., citation analysis, social network analysis, and h index) to Dr Porter, who are mostly involved into studies on research evaluation via term-based analytic models.

CONCLUSIONS

This paper proposed a diffusion-based network analytic methodology to predict the dynamics of scientific activities from two aspects: the potential research interest of scientific researchers and the future evolutionary directions of scientific topics. The method provides a novel way to integrate cutting-edge techniques in complex network analysis with term-based bibliometric approaches: 1) oriented to the unsupervised environments in real-world bibliometric studies, an indicator divergence is used to help automatically decide the number of topics for a K-means-based clustering approach; 2) science maps are introduced to bridge term-based bibliometric approaches (e.g., topic analysis, and term-based author correlation analysis) with a bipartite network, and provide supplementary information for further network analysis; and 3) a diffusion-based prediction approach is proposed to detect and predict the amount of resources diffusing within a bipartite network. The case study, applying the diffusion-based network analytic methodology to a set of bibliometrics-related articles, effectively identified a list of future evolutionary directions of specific bibliometrics-related topics and also a list of potential research interests of selected authors in the field of bibliometrics. The results can benefit not only bibliometric researchers for better understanding the research frontiers of bibliometrics but also a wide range of disciplines in information science and technology, science policy, and related industrial sectors.

We anticipate further study to look into the following directions: 1) introducing machine learning techniques to train our analytic models (in particular the prediction model) by dividing our dataset into several training sets with sequential time stamps and learning parameters (e.g., the diffusion strategy, and weights to indicate the priority of different terms) from the training process; 2) proposing more persuasive approaches for validation measures, e.g., comparing with the results of some similar bibliometric studies, expert knowledge-based evaluation, and quantitative validation models; and 3) testing the adaptability of our methodology by applying to diverse cases, e.g., relatively broad disciplines in natural science and social science (e.g., computer science, business and management), relatively specific areas (e.g., nanotechnology and solar cells).

ACKNOWLEDGMENTS

This work was partially supported by the Australian Research Council under Discovery Grant DP150101645 and the National Science Foundation of China under Grant 71673024.

REFERENCES

Abbasi, A., Hossain, L., & Leydesdorff, L. (2012). Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics, 6*(3), 403-412.

Börner, K. (2014). *Atlas of Knowledge*. Cambridge, MA: MIT Press.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., . . . Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS One, 6*(3), e18029.

Chen, C., Ibekwe SanJuan, F., & Hou, J. (2010). The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology, 61*(7), 1386-1409.

Ding, W., & Chen, C. (2014). Dynamic topic detection and tracking: A comparison of HDP, C-word, and cocitation methods. *Journal of the Association for Information Science and Technology, 65*(10), 2084-2097.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35-41.

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics, 51*(1), 69-115.

Huang, Y., Zhang, Y., Youtie, J., Porter, A. L., & Wang, X. (2016). How does national scientific funding support emerging interdisciplinary research: A comparison study of big data research in the US and China. *PLoS One, 11*(5), e0154509.

Hung, S.-W., & Wang, A.-P. (2010). Examining the small world phenomenon in the patent citation network: a case study of the radio frequency identification (RFID) network. *Scientometrics, 82*(1), 121-134.

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology, 68*(4), 984-998.

Leydesdorff, L., & Bornmann, L. (2012). Mapping (USPTO) patent data using overlays to Google Maps. *Journal of the American Society for Information Science and Technology, 63*(7), 1442-1458.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory, 37*(1), 145-151.

Noyons, E. C., Moed, H. F., & Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *Journal of the Association for Information Science and Technology, 50*(2), 115.

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science, 28*(6), 441-453.

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *nature, 435*(7043), 814-818.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science, 50*(9), 799-813.

van Eck, N., & Waltman, L. (2009). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538.

van Eck, N., Waltman, L., Noyons, E., & Buter, R. (2010). Automatic term identification for bibliometric mapping. *Scientometrics, 82*(3), 581-596.

Wang, X., Liu, Y., & Xiong, F. (2016). Improved personalized recommendation based on a similarity network. *Physica A: Statistical Mechanics and its Applications, 456*, 271-280.

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the Association for Information Science and Technology, 60*(10), 2107-2118.

Yan, E., Ding, Y., & Zhu, Q. (2010). Mapping library and information science in China: A coauthorship network analysis. *Scientometrics, 83*(1), 115-131.

Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change, 85*, 26-39.

Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology and a case study focusing on big data research. *Technological Forecasting and Social Change, 105*, 179-191.

Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017). Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics. *The Journal of the Association for Information Science and Technology*, to appear. doi: to appear

Zhou, T., Ren, J., Medo, M., & Zhang, Y.-C. (2007). Bipartite network projection and personal recommendation. *Physical Review E, 76*(4), 046115.

COPYRIGHT

81st Annual Meeting of the Association for Information Science & Technology | Vancouver, Canada | Nov. 9 - 14, 2018