OpenStyle Academic Hotpots Extracting Algorithm Based On Network Random Block Model

Xing Wang *
Renmin University of China, China, Beijing, China
e-mail: wangxing@ruc.edu.cn

Abstracts

Academic hot spots are required for many academic information management tasks including academic trends identification, funding policy decision and peer reviewing programming. Since the mid-1980s, network-assisted literatures as the carrier of knowledge and the spreading hot spots, its very nature has been a dramatic change from the paper style to the progressive of digital topic pieces resources with network. In this article, we assess the structure of hotspots from academic pieces of academic evidence. We also examine the emerging algorithm of weighted random block model for hotpots extraction purposes: unknown community detector on openstyle interdisciplinary parts so that they can be directly read and recovered from pieces text evidence. FN could provide a stable initial starts for Random block algorithm, and substantial further research is needed to make a thorough evaluation for its practical viability.

Keywords: Openstyle Hot spots, random block model, social network, pattern recognition.

1. Introduction

Academic hot spots are required for many academic information management tasks including academic trends identification, funding policy decision and peer reviewing programming. Driven by technology, literatures as the carrier of knowledge and the spreading tool of hot spots, its very nature has been a dramatic change from the paper style to the progressive of digital topic context resources and network resources. Literature becomes more and more easy to get, hot spots spread faster than before which enhance the knowledge recycling; Academic hotspots is also becomes more complexity like: the growing number of academic hotspots, areas of diversity, interdisciplinary has become increasingly evident. The raw data granularity is continued to refine which enhance the difficulty for new knowledge understanding. Interdisciplinary academic hotspots algorithm to track openstyle, hierarchical real-time hot topics which extract knowledge quickly, efficiently and reliably

This paper aims to study the algorithm for extracting academic hotspots, which is composed of 5 parts: the second part discusses the hot literature, the third part describes the algorithm, the fourth part gives empirical result, the fifth part us the conclusion.

2. The development of academic hotspots

The concept of academic hotspots was raised early as in 1965 by Sylvia Plath[1]. After nearly 60 years of continuous development and expansion, it has experienced the stages of the combination of primary literature and expert appraisal, determination of literature-coupled discipline structure, and content-based analytical construction of hotspots, specifically as follows:

(1)50-70s: combination of primary literature and expert appraisal.

Plath built the theory of using the literature references relationship to extract hotspots, the theory defined highly-cited literature as main subject literature, also known as the academic frontier, which is the basis for producing academic hotspots. Subject experts induce and conclude the academic hotspots according to the highly- cited literature[1]. In 1973, Henry Small inherited Plath's theory, and studied the citation relationship in the field of particle physics with co-citation-relationship analysis, creating the co- citation- relationship- based research on papers of academic hotspots[2] and defining the highly-cited literature as the academic hotspots. In 2009, Pan Jiaofen once clustered the co-cited literatures to create mutual relationship between with ESI (2002-2008) yearly literature reference data, and the used gravity model and visualization method to draw scientific activities panorama, producing 121 hot research fields.

(2)70-90s: determination of literature-coupled discipline structure.

With respect to the drawn material of academic hotspots papers, Weinberg held that it is not appropriate to define the academic hotspots as cited papers, because the cited papers usually

take a long publishing period [4]. Report from the 2009 humanities and social sciences journal evaluation implied that literatures on Humanities and Social Sciences have average half-life period of 4.2 years [5]. Conflict between citations quantity and timeliness cannot be solved, so the concluded themes from it reflect academic hotspots limitedly. Weinberger insisted that academic hotspots should be derived from highly-cited literatures, which can be obtained only from a large number of citations. In 1974, he put forward that the research focus is composed of the highly-cited literature citing articles (articles that cited highly-cited articles). In 1994, Persson inherited and developed Weinberger's theory, raising disciplinary structure theory based on literature coupling. Persson pointed that disciplinary structure, which is the basis of extraction of academic hotspots, consists of similar literature. Similarity in disciplinary structure can be represented by literature coupling, which is calculated by the number of similar reference in a couple of articles [6]. Compared with co-citation relationship, literature coupling measures the literature similarity more directly, without other supplementary information. For two given literatures, the coupling strength is fixed. However, the result based on co-citation varies as the number of literatures increase, and is vulnerable to disciplines half-life period and unbalance of literatures of different disciplines. In 1985, Luo Shisheng expanded the literature coupling to discipline coupling, specialized coupling and authors coupling[7]. In 2003, Morris, Yan and Wu introduced the timeline analysis on basis of literature coupling, to show the appearance and disappearance of the research focus. Literature strength calculated by the literature coupling is independent of other literatures and does not change with the emergence and disappearance of new literature. In 2012, Wang Lixue and Leng Fuhai pointed that the instantaneity makes literature coupling more appropriate for the study of academic hotspots than co-citation relationships [8].

(3)90s - Current, content-based analytical construction of hotspots.

With the integration of disciplines and research refinement, study on hotspots is extended from literatures to keywords of literature, which opening a new era of content-based construction of hotspots[9]. Co-word analysis is a representative method which is based on thesaurus. This method is also originated in the 1970s to the 1990s, and tends towards perfection after the constantly improvement done by Callon, Whittaker Courtial and other scholars. The core of co-word analysis is the co-occurrence matrix, which is used to measure the close correlation of thesaurus, and is indicated by the number of the co-occurrence of the two thesauruses. In 1991, Braam, Wed and Van, on the basis of co-word analysis research achievement, defined the research focus as: a set of research problems and concept that is under the common concern of scientific researchers with different knowledge and social background[10]. Co-word analysis describes disciplinary development with a set of thesauruses and explores academic hot spots.

Compared to the previous two methods, co-word analysis pays more attention to the content of the literature, leading to the formation of the disciplinary questions set composed of thesauruses and research focus method composed of high-frequency thesauruses. Data about co-word analysis is huge and data processing task is more arduous, especially thesaurus processing. Compared with the literature method, co-word analysis for the study of exploring academic hotspots is more direct, because academic hotspots is usually described by a word or phrase, which is correspond to the co-word analysis. Co-word analysis of academic study of hotspots is more refined, more comprehensive and reliable results can be obtained, and proceeding from thesaurus, we can grasp academic hotspots more accurately; timeliness is improved through the hotspots research based on thesaurus. In 2008, Zhong Weijin and Li Jia pointed out that under the continuous development of co-word, on one hand, thesaurus is transforming from indexing terms and thesaurus to free words gradually; on the other hand, co-word analysis develops from entire article to paragraphs and staements [11]. In 2011, Li Gang and WU Rui did the co-word analysis and academic hotspots mining analysis on published literature in the field of competitive intelligence during 2001-2010 in CNKI database. This article analyzes 3422 articles in the field of competitive intelligence during 10 years, and 6065 thesauruses is obtained, totally 16287 times in frequency. This is typical application to extract academic hotspots by co-word analysis method[12].

In summary, with the development of information technology, the literature base of academic

hotspots is changing from citations to citing articles sets, and the object of study is gradually transiting from literature analysis to content analysis. The extraction method improves towards more objective, timeless and robust, and the study basis of academic concept is built up on the core of thesaurus relationship, which is important in measuring the academic frontier.

The deficiencies of the current academic hot extraction is the existence of the performance of the content in the form of unity. Manifestations of the academic hot spots is the form of words set composed of high-frequency words in the literature, the lack of the necessary links between different words, an increase of academic hotspot explain the difficulty. Co-word matrix only to achieve a combination of academic hot elements, did not provide the link between the elements of the hot spots, which could result in the wrong academic hotspot explained, such as Financial Historical Survey of the political system, social relations, expressed as {financial, political, institutional, social relations, history}, then not accurate extraction academic hot spots, such as financial and social relations "financial history" financial system "will produce accurate statements is far different representation . Co-word matrix suitable for closure connectivity structure said, In fact, academic hot spots may not be restrained, but open. Closed community networks and open community network is shown in Figure 1 as shown in Figure 1, the imaginary circle represents the study found hot node represents the structural relationships of keywords, so generated research focus is very rich meaning, to reflect the relationship between the level of the concept, and can also be used to explain the possible connectivity between the hot spots and hot spots. This connectivity relations need to be estimated through the large number of documents, which reveals a stable level and the difference sequence of non-equilibrium hot structure.

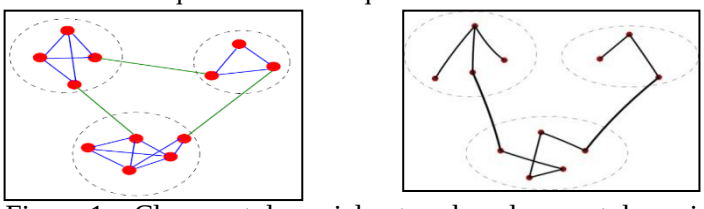

Figure 1：Closure style social network and open style social network

3. Random stochastic block model

(1) Standard stochastic block model

A stochastic blockmodel was first presented by SnijdersT.A.B, & NowickiK[14]. In this section we review briefly blockmodel, focusing on undirected networks since they are the most commonly studied. Let $G=(V,E)$ be an undirected multigraph on n vertices and m edges, possibly including self-edges, and let A_{ij} be an element of the adjacency matrix of the multigraph. Recall that the adjacency matrix for a multigraph is conventionally defined such that A_{ij} is equal to the number of edges between vertices i and j when $i=j$, but the diagonal element A_{ii} is equal to twice the number of selfedges from i to itself. Assuming $g=\{g_1,\ldots, g_k\}$ is the community, k is the community number; Let the number of edges between each pair of vertices be independently Poisson distributed and define ω_{rs} to be the expected value of the adjacency matrix element A_{ij} for vertices i and j lying in groups r and s, respectively. Now the probability of Graph G given the parameters and group design id as follows Eq(1):

$$P(G \mid g,\omega)=\prod_{i<j}\frac{(\omega_{g_ig_j})^{A_{ij}}}{A_{ij}!}\exp(-\omega_{g_ig_j})\prod_{i}\frac{(\frac{1}{2}\omega_{g_ig_i})^{A_{ii}/2}}{(\frac{A_{ii}}{2})!}\exp(-\frac{1}{2}\omega_{g_ig_i}) \tag{1}$$

Given that $A_{ij}=A_{ji}$ and $\omega_{rs}=\omega_{sr}$. Eq (2) can be rewritten in the more convenient form:

$$\log P(G|g)=\sum_{rs}\frac{m_{rs}}{m}\log(\frac{m_{rs}/m}{n_rn_s/n^2})=\sum_{rs}p_k(r,s)\log\frac{p_k(r,s)}{p_1(r,s)} \tag{2}$$

(2). Degree corrected stochastic block model

In degree-Corrected block model presented by Karrer & Newman[15], let the expected value of the adjacency matrix element A_{ij} be $\theta_i\theta_i\omega_{gigj}$. Then graph G has probability as Eq (3)

$$P\left(G|g,\omega\right)=\prod_{i<j}\frac{(\theta_i\theta_j\omega_{g_ig_j})^{A_{ij}}\exp(-\theta_i\theta_j\omega_{g_ig_j})}{A_{ij}!}\prod_i\frac{(\frac{1}{2}\theta_i^2\omega_{g_ig_i})^{A_{ii}/2}\exp(-\frac{1}{2}\theta_i^2\omega_{g_ig_i})}{\frac{A_{ii}}{2}!} \quad (3)$$

In modeling Randomized block model and corrected randomized block model, the given block number is the starting point of the algorithm. A general idea is to use the bootstrap for the community simulation, and then compare the different community corresponding to the network maximum likelihood estimation value, select the maximum likelihood value corresponding to the community as estimates of the true class. However, this strategy will cause over fitting without additional constraints for the likelihood function; the second idea is FN proposed by Newman M. (2004) [16] , like Eq (4) describe, first to estimate the community by maximizing network module. The module aims to measure the degree of network to maximize the model, improve the result based on the initial value of the randomized block model,

$$Q=\frac{1}{2m}\sum_{vw}\left(A_{vw}-P_{vw}\right)\delta\left(C_v,\ C_w\right) \quad (4)$$

A_{vw} is the number of edges between vertices v and w in the diagram. A_{vw} values of 0 or the weight of c between nodes relationship or strength of pairwise relations. m is the total number of edges, the constant term 1/2 is a normalization coefficient. In hotpots problem, weights between edges can be interval value between 1 and 0, so we modified we Q to be as Eq.(5).

$$\begin{aligned}Q&=\frac{1}{2W}\sum_{vw}\left(W_{vw}-\frac{s_vs_w}{2m}\right)\delta\left(C_v,\ C_w\right)\\&=\sum_i\left(e_{ii}-a_i^2\right)。\end{aligned} \quad (5)$$

Here we use the weighted graph by Karger in 1993 which to extend the weight {0,1} to an integer weighting. We named this algorithm as WFN -- to find in weighted graphs on the basis of FN algorithm. Here we use weighted WFN algorithm used for the initial classification of the randomized block model, initialization modeling data, and then into the correction randomized block model solution, the first use of the WFN initialize the network structure, and then randomized block model, as follows two-stage modeling algorithm:

Two Stage Modeling:

Step1: To apply WFN find the initial starting community.

Step2: Using Kernighan-Lin steepest descent algorithm in searching the most stable result by Bootstrap communities generated by the evaluation of the stability of results.

4. Empirical Study

Raw Data is between March 2011 recorded from the document downloading from a typical University Library scholars and their usage login to. Select the number of downloaded more than 10 literature by cleaning effective literature 3001, to extract the article keywords to generate 8959 keywords, as amended randomized block model can be estimated from the network structure more obvious block structure, to extract the 6 academic hot spots which have a high consistent with released report in 2011 by The Information Center in Social Science of RUC respectively. For summarize, we only list the two important group (1) the real estate market and local government bonds; (2) the level of information and capability developments by information level-building.

(1) Hotspots 1：Estate and local government Bonds

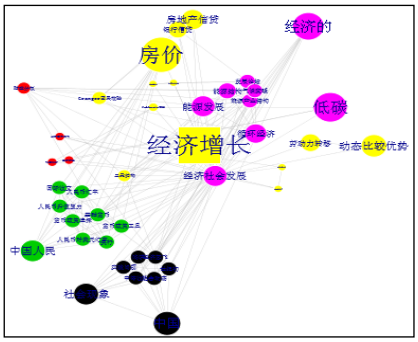

Figure 2 Estate and local government Bonds

Figure 2 shows that "economic growth" nodes play the role of connecting four academic hot. These hot spots were involved in recycling economy, housing prices, energy development, middle-income, monetary policy and fiscal policy. Academic economic growth is very comprehensive and in-depth, and economic growth as an integrated node, has a strong interdisciplinary characteristics.

(2) Hotspots 2： Government service capability developments by information level increasing

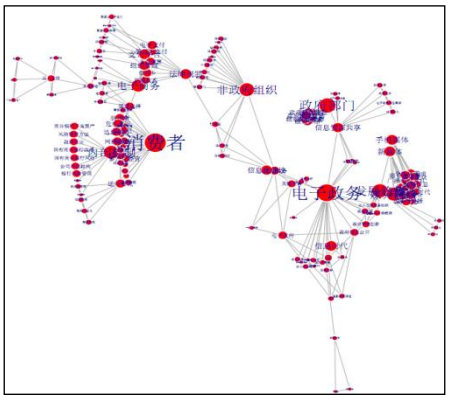

Figure 3 Government service capability developments by information level increasing

From Figure 3, the core is"consumer", "e-government" and "Government departments. Directly from these three hot core, the traditional methods of its almost know nothing about the internal structure of the hot spots, not to reveal the hot internal logic and laws. Correction randomized block model results to see, "e-government" is to strengthen the government management innovation "and" Open Government Information "and" trends "in the" information age "by government departments" and "information resources sharing" society "consumers" and other non-governmental organizations "to provide" public service ". Therefore, summarize and extract the the hot main content for government services, capacity-building, so that a high-level overview of the main content of the hot, traditional methods can not be achieved.

5. Conclusions

For academic hot extraction algorithm, try to establish based scholars focusing on literature analysis algorithms refined literature academic hot, the algorithm by the underlying data structure extraction and visualization of three aspects of composition. Introduced in the data underlying design scholars Readings data rather than literature published data, fully reflects the social impact of the academic hot spots; randomized block model-based academic hot extraction model is designed to avoid direct use of the randomized block model may lead to The block structure is unstable, we propose a the WFN-based randomized block model as one of the two-stage method, simulation and empirical studies show that the model can achieve better results in the hot extraction.

The Framework of the proposed algorithm based on the article, Renmin University of China scholars study sample, to extract the six academic hot spots, namely (1) the real estate market and local government bonds; (2) the financial crisis and climate change; (3) China's economy sustainable development; (4) the level of information and capacity-building; (5) market

economy and modernization of agriculture; (6) new media trends. Hot spots 4, 5, 6 and Renmin University of China Data Center released "2011 China top ten academic hot spots" study compared achieved initial consistency, confirmed the conclusions of the effectiveness and feasibility of the model.

Although empirical research to achieve better results, but still need the efficiency of algorithms on statistical design, the initial point do continue to study the stability of the algorithm research; conclusion of the current article is mainly humanities and social sciences disciplines scholars perspective derived conclusions based scholars in science and engineering disciplines, further validation; lack of a comparison of the different perspective of data, in particular, citing data and Selected Readings data mechanism research remains to be to continue to explore; the The algorithm also affected by the impact of disciplines scale, sub-disciplines of law and give a reasonable evaluation.

Hotspot detection based on large-scale academic literature data, the model and algorithm research has just started, the article is mainly based on the methods and models of scholars perspective, firm conclusions should be finished combining comprehensive and objective and expert's opinion which aims to understand the academic hot spots to provide more reliable information more easily accessible academic services for the majority of scholars.

References

[1] Morris SA, Yen G, & Wu Z. Time Line Visualization of Research Fronts[J].Journal of American Society for Information Science, 2003.54:413-422.

[2] Henry Small. Co-citation in the Scientific Literature：A New Measure of the Relationship between Two Documents[J].Essays of an information Scientists, 1973.2：256-269.

[3] Panjiaofeng. Scientific Structure Map [M]. 2009. Beijing: Science Publication (Chinese Version).

[4] Wenberg. Bibliographic coupling：a review. Information Storage and Retrieval[J], 1974.10：189-196.

[5]Su XinNing. Humanities and Social Science academic influence report (2009), 2010 China Social Science Press (Chinese Version).

[6] Persson. The intellectual base and research fronts of JASIS 1986-1990[J].Journal of the American Society for Information Science, 1994.45:31-38.

[7]Luo Shisheng. Coupling Type and Analysis [J].Library and Information Science,1985.1: 42-47(Chinese Version).

[8]Wang Lixue,Leng Fuhai. Frontier and its bibliometric identification method [J].Theory and exploration, 2010.33: 54-58. (Chinese Version)

[9]LiaoShengjiao,Xiaoxiantao. Research Advances on the bibiometrics - based Co - word Analysis[J],Information Science, 2008,06. (Chinese Version)

[10] Braam R, Wedh, & Vanraan. Mapping of science by combined co-citation and word analysis II dynamical aspects[J]. Journal of American Society for Information Science.1991.42(4):252-264.

[11]Ligang,Wurui, Research hotspot in the field of national nearly a decade of competitive intelligence analysis - based on analysis of common words[J]. Information Sciences (Chinese Version), 2011.29:1290-1293.

[12]Liu dachun. Humanities and Social Sciences Research Evaluation System [M],2009, Beijng, Econometric Science Publication.

[13]Talja,S.,Vakkari, P., Fry, J. & Wouters,P.The impact of research cultures on the use of digital library resources, Journal of the American Society for Information Science and Technology[J],2007.58：1674–1685.

[14]SnijdersT.A.B, & NowickiK. Estimation and prediction for stochastic blockmodels for graphs with latent block structure[J]. Classification.1997.14:75-100.

[15]KarrerBrian, & NewmanM.E.J. Stochastic blockmodels and community structure in networks[J]. American Physical Society, 2011.16:1-10.

[16]M.E.J.Newman. Fast algorithm for detecting community structure in networks [J].PHYSICAL REVIEW, 2004.69(6). 066133.