Scientometrics (2014) 101:1253–1271
DOI 10.1007/s11192-014-1347-y

Analyzing evolution of research topics with NEViewer: a new method based on dynamic co-word networks

Xiaoguang Wang · Qikai Cheng · Wei Lu

Received: 17 November 2013 / Published online: 22 June 2014

Abstract Understanding the evolution of research topics is crucial to detect emerging trends in science. This paper proposes a new approach and a framework to discover the evolution of topics based on dynamic co-word networks and communities within them. The NEViewer software was developed according to this approach and framework, as compared to the existing studies and science mapping software tools, our work is innovative in three aspects: (a) the design of a longitudinal framework based on the dynamics of co-word communities; (b) it proposes a community labelling algorithm and community evolution verification algorithms; (c) and visualizes the evolution of topics at the macro and micro level respectively using alluvial diagrams and coloring networks. A case study in computer science and a careful assessment was implemented and demonstrating that the new method and the software NEViewer is feasible and effective.

Keywords Science mapping · Co-word analysis · Network communities · Topic evolution · Emerging trend detection

Introduction

Since the 1950s, big science has rapidly developed (Price and de Solla 1963). With the flourishing of science, recognizing and grasping scientific frontiers and research trends in a timely manner has become more important and difficult than ever for scholars and science policymakers (Goth 2012). In face of this demand, some scholars in information science

X. Wang (✉) · Q. Cheng · W. Lu
School of Information Management, Wuhan University, Wuhan, China
e-mail: wxguang@whu.edu.cn

X. Wang · W. Lu
Center for Information Resources Research, Wuhan University, Wuhan, China

have focused attention on Emerging Trend Detection (Pottenger and Yang 2001; Roy et al. 2002; Kontostathis et al. 2003, 2004; Le et al. 2005).

Supporting disclosure of evolving and emerging trends in science, full-text paper databases, citation databases, abstract databases and patent databases, have gradually made this effort more feasible. Several applications, such as ThemeRiver, Bibexcel, CiteSpace II, Network Workbench, VOSviewer, and SciMAT, have been developed and have been used to achieve advances in the area of Scientometrics and Informetrics (Cobo et al. 2011a). Science maps drawn by these applications display the cognitive structure and dynamics of a research field (Börner et al. 2003).

In this paper, we propose a set of new methods based on co-word networks and complex network theory to reveal the evolution process of topics in a research field. Software, called NEViewer, was also developed based on the proposed methods. As a case study, the thematic evolution of computer science field was analyzed by only considering the papers published in five conference proceedings.

This paper is organized as follows. "Background" section gives a brief overview of the related research. "Research design" section introduces the proposed approach to analyze the evolution of a research field and the software NEViewer. "Case study" section displays the results of a case study. Conclusions and shortcomings of our research are drawn and discussed in "Discussion" section.

Background

Emerging trend detection and topic evolution analysis

In 2003, Kontostathis put forward the emerging trend concept for subject areas arousing the interest and discussion of more and more scholars (Kontostathis et al. 2003). Emerging trend detection (ETD) means recognizing emerging topics and their correlations in a scientific field. Various types of techniques have been developed to detect research trends; the most commonly used being citation-based analysis and keyword-based analysis. Since citation-based analysis takes longer and is not as clear as keyword-based analysis for the representation of emerging topics, researchers are prone to use the latter.

ETD can be divided into three phases (Le et al. 2005): topic representation, topic identification and topic verification. In the first phase, a large-scale literature dataset is collected and used to extract research topics. The key to this phase is to normalize keywords that represent a concept semantically. To improve the effectiveness of word sense disambiguation and the preciseness of topic convergence, various types of machine learning technologies have been introduced by researchers, such as Neural Network algorithm and Latent Semantic Indexing (LSI). They has been used to identify bursting topics (Pottenger and Yang 2001) and cluster concepts (Kontostathis et al. 2004). As the meaning of Single Value Decomposition is indefinite in LSI, it is hard to control the effect of topic clustering and further, the algorithm complexity is high. van Eck et al. (2010) applied probability latent semantic analysis (PLSA) to identify topics from in a corpus of scientific literatures; PLSA provides a kind of fuzzy clustering of the linguistic units occurring in a corpus and reduces time complexity with a better semantic distinction effect than LSI (Hofmann 1999). Each cluster corresponds with a topic.

In the phase of topic identification, researchers usually discover research topics with similarity matrix and hierarchical clustering (van Eck and Waltman 2009; Klavans and Boyack 2006). Recently, complex network theory has gradually been used in this phase,

which has promoted knowledge network analysis as a burgeoning approach in ETD (Amitay et al. 2004; Chen and Redner 2010). In 2009, Wallace et al. researched on clustering in co-citation network and proved the natural advantages of community finding algorithms in topic detection. In 2010, Chen and Redner (2010) analyzed the citation network data for over 100 years from the *Physical Review* series, revealing the corresponding relationship between citation network communities and the evolution process for them.

In the third phase, the traditional method, keyword frequency, is still popular (Buente and Robbin 2008), but researchers have started to make use of more complex methods to verify emerging topics. Le et al. (2005) put forward a method to evaluate popularity and availability according to six features of research topics. Lee et al. (2010) utilized data profile and paralleled adjacent clustering algorithm in measuring three developing phases and features of digital library research. Schiebel et al. (2010) studied terminology diffusion model, and methods for exploring old and new topics and their correlation structure by diachronic cluster analysis. Tu and Seng (2012) proposed two new indexes, New Index (NI) and Published Volume Index (PVI), to decide the emerging topics. Chavalarias and Cointet (2013) showed the phylomemetic patterns in science evolution by analyzing some sequential structural properties of scientific fields.

In summary, detecting the emerging trend by mixing of keywords, text mining technologies and information visualization technologies has become more prevalent than ever. With the development of science mapping research, methods to reveal the developing state of research topics from the aspect of network dynamics is now evolving into a new path (Herrera et al. 2010; Liu et al. 2013). From the general experiences of science activities, we know that time dimension should be considered in the process of verifying whether a topic is an emerging one, a hot one, a obsolete one or a dead one. It means when considering the trend of a special topic, its past performance should also be considered carefully. Research topics always arise from some knowledge bases (Chen 2005). Therefore, to predict a trend in a research field, it is essential to recognize the evolution and life cycle of all topics in this field, and is of great significance when detecting emerging topics and forecasting trend dynamics.

Co-work network and network community

Leydesdorff, Boyack, Börner, Chen et al. have made remarkable achievements in the field of knowledge network analysis and science mapping (Leydesdorff and Rafols 2008; Boyack et al. 2005; Börner et al. 2003; Chen 2005). Their work showed the significance of knowledge networks on latent knowledge discovery, recognizing research frontiers, and ETD (Mane and Börner 2004). A Co-word network is a kind of knowledge network usually constructed with author keywords and their co-occurrence. Previous research has shown that co-word networks can not only serve for science mapping in a specific field, but are of methodological significance in other knowledge networks. Compared with the traditional Bibliometrics methods based on citation or co-citation, co-word network has more advantages in terms of timeliness (Pottenger and Yang 2001; Roy et al. 2002).

Community, a common phenomenon in networks, is aggregated by a group of highly intensive and closely related nodes in a network. There is a high density inside a community and a low one between different communities. In a real network, nodes belonging to the same community are more likely to share similar attributes and functions. For example, WWW, web pages belonging to the same community usually have similar themes; in a scientific collaboration network, scholars share similar interests from multifarious research communities at different levels named invisible college or subjects.

As a mesoscopic phenomenon, community is conducive to the cognition of relationships between network structures and their functions, so attracts more and more scholars' attention. Newman and Girvan's (2004) works indicated that there exist obvious communities in citation and co-author networks. Boyack et al. (2005) also revealed clusters in global science map. The prevalent existences of communities in knowledge networks indicate their essentiality. Lambiotte and Panzarasa (2009) held that communities, closely related with disciplines and subjects, can be seen as a territorial partition mechanism for science mapping and landmarks for research frontiers. In addition, communities in knowledge networks are of great significance to knowledge creation and distribution at different level. Wang (2013) have found that communities in social network contribute to the reduction of collaboration costs for participants as well as discovering and transferring knowledge.

Network community detection

Many community detection algorithms have been proposed by computer scientists and physicists in the last decade. There are two main approaches to identify the latent structures within a given dataset: the network topology based approach and the content-based approach (Ding 2011). The first approach is based on *Graph Theory*. Modularity Maximization is a widely adopted method for community detection (Newman and Girvan 2004). Modularity is designed to measure the strength of divisions of a network into communities. The high modularity implies networks have dense connections within communities but sparse connection between communities.

Although modularity was designed initially for the unweighted and undirected networks, it has been extended to the weighted and directed networks. However, it has been shown that modularity is unable to detect small communities. In addition, the Modularity Maximization method is weak on detecting overlapping communities. For this reason, Palla et al. (2005) proposed a K-cliques algorithm, in which each node is able to be assigned to multiple communities. Ball et al. (2011) built an algorithm for finding overlapping communities and improved the ability to discover either overlapping or non-overlapping communities.

In 2008, McCain (2008) adopted the second approach and found the effectiveness of content-based topical analysis in citation networks. Later, Wallace et al. (2009) found that the technique developed by Blondel et al. (2008) is robust and efficient and that the results generated can be of great use to study various facets of the structure and evolution of science.

Research design

Scientific literatures are important carriers for scholars to publish their research outcomes (Cobo et al. 2012; Wang et al. 2010). Usually, the authors are required to provide several carefully selected keywords to represent the main research topic of a paper. A co-word network is constructed based on keywords in a set of document. By adding temporal and longitudinal information to the co-word networks, researchers are able to map the evolution of a research field (Garfield 1994; Wang et al. 2010; Cobo et al. 2011b).

Based on the above idea, we propose a new approach in a longitudinal framework for the evolution analysis, which is divided into four phases: topic representation, topic identification, topic evolution analysis and visualization. First, we preprocessed the raw

data and convert them into a sequence of temporal co-word networks. We assigned one time stamp for each of those networks based on the publication date. Then, community detection algorithms were adopted to uncover the latent community structures within the co-word networks in each time stamp. Each community is assumed to be a corresponding topic. After that, we determined the incidence relations among different communities. Meanwhile, one or more representative keywords were selected to label each community as the corresponding topic. The last step was to visualize the evolution procedures. The new research framework and workflow is depicted in Fig. 1.

Construction of co-word network

To construct co-word network, we first define as follows: a sequential document set as $D = \{D_1, D_2, \ldots, D_n\}$, where D_t is a collection of documents published during the period of t, $D_t = \{d_{t1}, d_{t2}, \ldots, d_{tm}\}$, d_{tm} is the document numbered m during the period of t, $W = \{w_1, w_2, \ldots, w_n\}$, w_k is the keyword numbered k in a document d; co-word networks $G = \{G_1, G_2, \ldots, G_n\}$, G_t is the co-word network during the period of t. $G_t = \{V, E\}$, V is the node set and E is the relationship set.

The definition of relationships in the co-word networks is as follows:

1. Word w_a and w_b, if $w_a \in d$ and $w_b \in d$, then w_a and w_b have a co-occurrence relationship for one time, here relationship is not weighted.
2. If w_a and w_b co-occurrence in n documents, then there is a connection with weight n between w_a and w_b.

According to this rule, given D_t, the construction of network G_t is as follows:

1. Construct a empty co-word network G_t;
2. Traverse documents in document set D. For every document d, its conception descriptive word W. For each w, if w does not appear in G, add w into G as a node; for any word combination $w_a w_b$ in W: if there is no connection between w_a and w_b in G, then build a link between them and set relationship weight to 1; if there is a connection already, 1 is added to the link weight.

To obtain sequential co-word networks, we need to divide document sets according to time slice. The TimeLine method and fixed time window are two commonly used methods (Sun et al. 2007). The TimeLine method is very complicated and can not ensure an

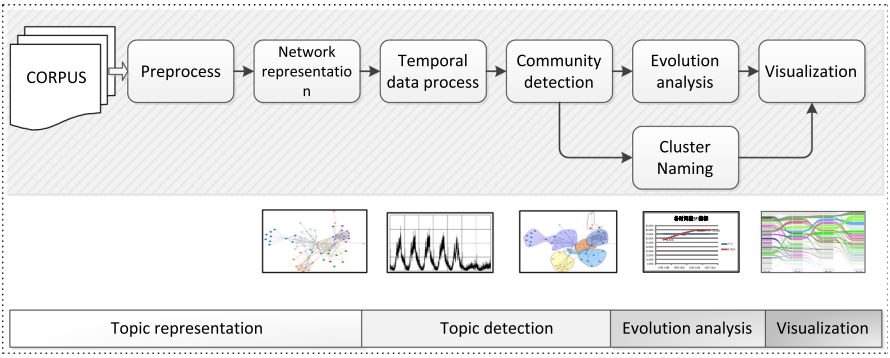

Fig. 1 The workflow of our research framework

effective division. So this paper applies the fixed time window method: set a time period with length t as one time window, divide document sets into n parts, then construct co-word network for each part respectively, and obtain co-word network sequence G.

Usually, in the workflow of science mapping, normalization and similarity measures will be used over the data (association strength, Equivalence Index, Inclusion Index, Jaccard Index, and Salton's cosine) when the network of relationships between the selected units of analysis has been built, and following by the clustering of data. In our research framework, there is no need to perform normalization and similarity measures as we do not process data with dimensionality reduction techniques such as principal component analysis or multidimensional scaling, or clustering algorithms. We discover knowledge directly from networks with community detection technologies as discussed in the next section.

Communities detecting and corresponding topic labeling

Previous research has showed that there exist communities in co-word network which are similar with clusters from the perspective of clustering analysis in data mining. Communities at different scales represent various granularity of a research field. Lancichinetti and Fortunato (2009) have compared several methods and found that three algorithms introduced by Rosvall and Bergstrom (2007, 2008), Blondel et al. (2008) and Ronhovde and Nussinov (2009) have an excellent performance for community detection.

After community detection, the next step is to find and label the corresponding topic for each community. Since the basic nodes of co-word network are keywords, to find and label the corresponding topic for each community means to find one or more core nodes inside each community.

There are several indexes for ranking nodes in a network, such as *Centralization*, *Prestige* and *PageRank* etc. (Costa et al. 2007). These indexes all rank nodes on the global level other than regional level, so they do not match our request for ranking nodes in a specific community. Therefore, we used the within-module degree z-score, a index, proposed by Guimerà et al. (2006) to evaluate node. z-score can rank nodes well from regional level other than global level.

$$z_i = \frac{k_{s_i}^i - \left\langle k_{s_i}^j \right\rangle_{j \in s_i}}{\sqrt{\left\langle (k_{s_i}^j)^2 \right\rangle_{j \in s_i} - \left\langle k_{s_i}^j \right\rangle_{j \in s_i}^2}}$$

k_s^i is connection number of other nodes from node i to community s, s_i is the community where node i belongs, and $\langle \cdots \rangle_{j \in S}$ is average number of all nodes in community s. The higher z-score, the closer the relationship is between nodes and other nodes in the same community. The within-module degree z-score reveals the representative nodes in a special community, namely corresponding topic. According to Guimerà's study, the nodes that have a high within-module degree ($z \geq 2.5$) are hubs and representative ones.

Community evolution determining

From the perspective of knowledge sociology, science knowledge are created synergistically. Old knowledge is the base of new knowledge. In a specific research field, social

factors such as recognized basic theories, spreading theory models and research genres will impact, oblique or direct, hereditary characters of knowledge creation e.g. ideas, themes, schools and theories. However, with a changing social environment, recruiting and retiring of scholars, discovery of new scientific phenomenon, new topics emerge and interact continually. Usually, the new topics and old ones are interconnected, which leads to the evolution of a research field.

Once the research topics are indicated by network communities, the revolution analysis of a research field turns into the analysis of dynamics of communities. Following the proposition raised by Palla et al. (2007), the evolution of communities can be divided into six forms: *Birth*, *Growth*, *Merging*, *Contraction*, *Splitting* and *Death*.

D1: *Birth*: communities that do not exist in the time period of t, and emerge during the time period of $t + 1$;
D2: *Growth*: communities exist in the time period of t, and will exist in the time period of $t + 1$ with a larger scale;
D3: *Merging*: two or more communities in the time period of t, and merge into a new community in the time period of $t + 1$;
D4: *Contraction*: communities exist in the time period of t, and will exist in the time period of $t + 1$ with a smaller scale;
D5: *Splitting*: communities in the time period of t and are split into two or more new communities in the time period of $t + 1$;
D6: *Death*: communities that exist in the time period of t, and do not exist during the time period of $t + 1$;

Each form of evolution requires analyzing the community structure in both time stamp t and time stamp $t + 1$. In our research, we simplify the evolution analysis as finding the appropriate successors and predecessors in two sequent time windows except the *Birth* and *Death*, in which there is no predecessors and successors respectively.

For finding predecessors and successors of communities, the similarities of different communities should be measured. We assume that if the similarity between two communities in two sequent time windows is larger than a certain threshold, the evolution relationship exists.

We define the predecessor of community $M_{(t+1)j}$ as $Pre(M_{(t+1)j})$ with the following formula, in which δ is the threshold value of similarity and *Sim* measures the similarity.

$$pre(M_{(t+1)j}) = \{M_t | M_t \in G_t, Sim(M_t, M_{(t+1)j}) > \delta\} \quad (1)$$

The similarity function *Sim* is the pivotal in the community evolution analysis. There are three basic methods to measure community similarity: node-based measurement (Palla et al. 2007), relation-based measurement (Berger-Wolf and Saia 2006) or a mixed one.

Node-based community similarity measuring

The basic measurement indexes of node-based method include *Dot Product*, *Cosine*, *Jaccard Coefficient* and *Generalized Jaccard Coefficient* etc. A new weighted matching index is given here. Given community M_x and M_y, their corresponding word sets are C_x and C_y, the definition of weighted matching index is:

$$Sim(M_x, M_y) = \frac{\sum_{v \in C_x \cap C_y} W(v)}{\min\left(\sum_{v \in C_x} W(v), \sum_{v \in C_y} W(v)\right)} \quad (2)$$

$W(v)$ is the frequency of node v, and $min(x,y)$ is the smaller value of x and y.

If it is required that two communities are not only similar in nodes but also in node scale, to use *Dot Product* is a good choice. However, in topic evolution analysis, there may be a huge difference in node scale between communities. To allow the existence of this situation, we can use *Cosine* or *Generalized Jaccard Coefficient*. To use *Jaccard Coefficient* in vector of dual attribute data is a simpler way. Weighted similarity, the steadiest index, can be used to measure the similarity between communities under most circumstances.

Core node-based community similarity measuring

Usually, the development of community is mainly dependent on the core nodes. Core nodes are the important ones in a community. In our study, we use z-score to measure the importance of nodes in a community. If a node's within-module degree z-score ≥ 2.5, the node is treat as a core node. We use $H(M_x)$ to represent the collection of core nodes in a community M_x. Then the similarity of two communities, $HS(M_x, M_y)$, can be defined as follows:

$$HS(M_x, M_y) = sim(H(M_x), H(M_y)) \quad (3)$$

Here, *sim* function can be *Cosine*, *Generalized Jaccard Coefficient*, *Dot Product*, etc.

Relationship-based community similarity measuring

Berger-Wolf and Saia (2006) has proposed a relationship-based algorithm to measure the similarity of two communities. Based on this idea, Wu et al. (2010) put forward a simple equation, which measures the similarity of two communities M_x and M_y.

$$ES(M_x, M_y) = \frac{E(x) \cap E(y)}{E(x) \cup E(y)} \quad (4)$$

In the formula (4), $E(x)$ represents the collection of edges in a community M_x. $ES(Mx,M_y)$ is the similarity of two communities. $|E(x) \cap E(y)|$ is the size of common edges in M_x and M_y. $|E(x) \cup E(y)|$ is the size of edges in the union of M_x and M_y.

Through the above three ways and a default threshold value, we can find the predecessors and successors for any community at any time window.

Community evolution visualization

Visualizing the process of community evolution is vital to understanding the dynamic of research field. Rosvall and Bergstrom (2010) applied alluvial diagram originated from geography to map the evolution of network. Figure 2 provides such an example of alluvial diagram. The colored rectangle areas in Fig. 2 represent communities and their sizes; the colored curve areas between two time stamps denote the evolution process. If one colored rectangle area in time stamp t divides into two same colored areas in time stamp $t + 1$, it implies that one community divides into two communities; if two colored rectangle areas

in time stamp t merges into the same colored area in time stamp $t + 1$, it implies that two communities combine to form a large community, or a new community is created.

We adopted alluvial diagram to illustrate the evolution of research topics. However, its shortcomings are still obvious. Firstly, the original alluvial diagram can not reflect position of each topic during a time window; secondly, if there are two or more predecessors for one community, the importance of each predecessor in the merging process can not be shown.

To solve the disadvantages, some ameliorations are made to Rosvall's method. At first, each community is ranked in two ways: one is based on the sizes of communities; the other is based on the degree centralities of communities. Secondly, a coloring network diagram is introduced to visualize each community. This kind of diagram is designed to reveal more details of the community evolution process by providing the successors and predecessors for each node.

There are two types of coloring algorithms: the *Forward Coloring* algorithm and the *Backward Coloring* algorithm. The forward coloring helps uncover the trend of each node in a community and the backward coloring helps reveal the source of each node in a community. For example, there is one community *A* in the time stamp *Time1* (see Fig. 3) and two communities *B* and *C* in the time stamp *Time2*. Communities *B* and *C* are the successors of community *A*. The *Forward Coloring* algorithm will assign different colors for the nodes in community *A* based on their divisions in time stamp *Time 2*. Assumed that in the future in the time stamp *Time 3*, community *B* and community *C* merges into one large community *D*. The *Backward Coloring* algorithm will assign different colors for nodes in community *D* based on their sources in time stamp *Time 2*.

We formalize the rules of *Forward Coloring* and *backward coloring* as follows: (1) *Forward Coloring* algorithm: given the community M_t in time stamp t, for any node (keyword) v in this community, if the same keyword v also occurred in community $sM_{t+1,i}$, we let $VColor(v) = AColor(sM_{t+1,i})$, in which $sM_{t+1,i}$ represents the successor community of community M_t. AColor(M) represents the color of community M_t in the alluvial diagram, VColor(v) denotes the color of node v in community M_t. (2) *Backward Coloring* algorithm: given the community M_{t+1} in time stamp $t + 1$, for any node (keyword) v in this community, if the same keyword v also occurred in community $pM_{t,i}$, we let $VColor(v) = AColor(pM_{t,i})$, in which $pM_{t,i}$ represents the predecessors community of community M_t.

Further, we also adopt a hierarchical layout algorithm to the coloring network diagram, put the core nodes in the centre area and enlarge their sizes. We want to emphasize and show the importance of the core nodes in a community by the layout algorithm.

NEViewer

We developed software, in Java, based on the above framework and methods and named it NEViewer (Network Evolution Viewer), which supports the NWB file format used by Network Workbench (http://nwb.cns.iu.edu/).

We realized all of the algorithms mentioned above in NEViewer. Researchers can choose several algorithms to detect communities, measure the similarities of communities, rank nodes in a community or communities in a network slice, and show the details of a community by coloring a network diagram. Some basic complex network metrics are also supported in this software such as the PageRank score and Centrality Degree.

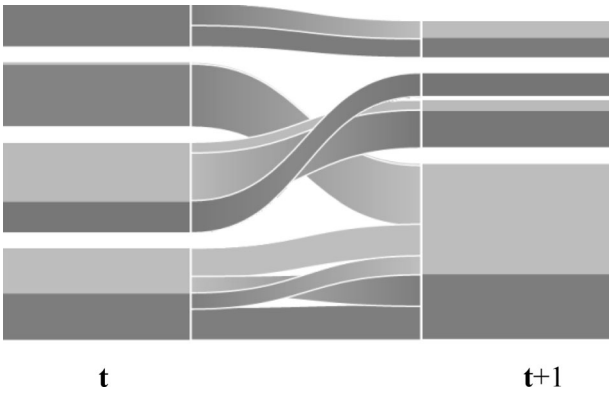

Fig. 2 An example of alluvial diagram

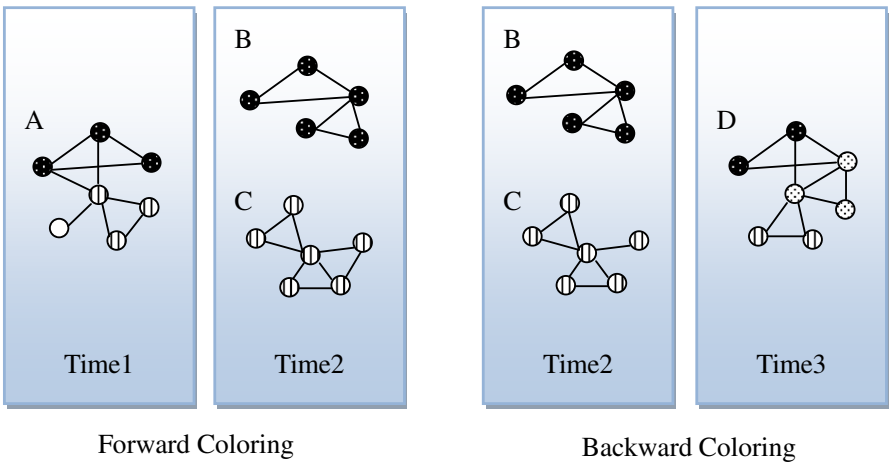

Fig. 3 The backboard coloring network and forward coloring network

Case study

Dataset

In order to evaluate the effectiveness of our method, we conducted a case study using NEViewer. Papers from five conferences were collected, comprising the five main conferences in the field of *Information Retrieval*, *Data Mining* and *WWW* (i.e. KDD, SIGIR, CIKM, CSCW and JCDL). The dataset contains 7,234 papers published from 2000 to 2011. Workshop papers were excluded from the dataset. Each paper in the dataset includes a title and an abstract. Both stemming and stop words are removed.

Table 1 Basic network attributes for three datasets

Measurements	Datasets		
	T1	T2	T3
Nodes	1,217	2,295	2,903
Isolated nodes	9	16	17
Edges	12,076	25,678	33,272
Mean degree	19.84	22.37	22.92
Density	0.00816	0.00488	0.00395

Sequential network construction

The dataset is divided into three periods according to paper's publish year: T1 = [t_{2000}, t_{2003}], T2 = [t_{2004}, t_{2007}] and T3 = [t_{2008}, t_{2011}]. 2,480 papers are included in T1, 4,283 papers in T2 and 5,517 papers in T3. Three co-word networks N1, N2 and N3 are constructed. In total, 1,217 nodes (keywords) and 12,076 edges appear in N1, 2,295 nodes 25,678 edges in N2, and 2,903 nodes 33,272 edges in N3. The basic network attributes are shown in Table 1.

After constructing the sequential network, Blondel's algorithm is applied on the datasets. The results are shown in Table 2.

Topic evolution analyzing

To analyze and visualize the overall topic evolution, we adopted the node-based similarity measuring algorithm and alluvial diagram mentioned above. Figure 4 gives a global picture of the topic evolutions from 2000 to 2011, in which we ignore the communities including 10 nodes or less because they usually do not have palpable influences on the whole context. The global evolution discovers insights of the major topics related to Information Retrieval, Data Mining, and the World Wide Web.

As shown in Fig. 4, the alluvial diagram displays different types of evolution mentioned by Palla et al. (2007). The *Information Retrieval* community in 2004–2007 splits into several small communities in 2008–2011. The *Interaction Design* community and partially *Education* community in 2000–2003 merges into a big community *Interaction Design* in 2004–2007. *Multi-touch* births in 2008–2011, *XML* growths in 2004–2007, *Navigation* contracts in 2004–2007, and *Digital Library* dead in 2000–2003.

In terms of network position, most of the continuous topics fluctuate during the three time windows. That means research communities always change their research focuses at

Table 2 Communities in three datasets

Dataset	Number of communities	Average number of nodes in each community
T1	23	50.71
T2	34	65.57
T3	41	69.12

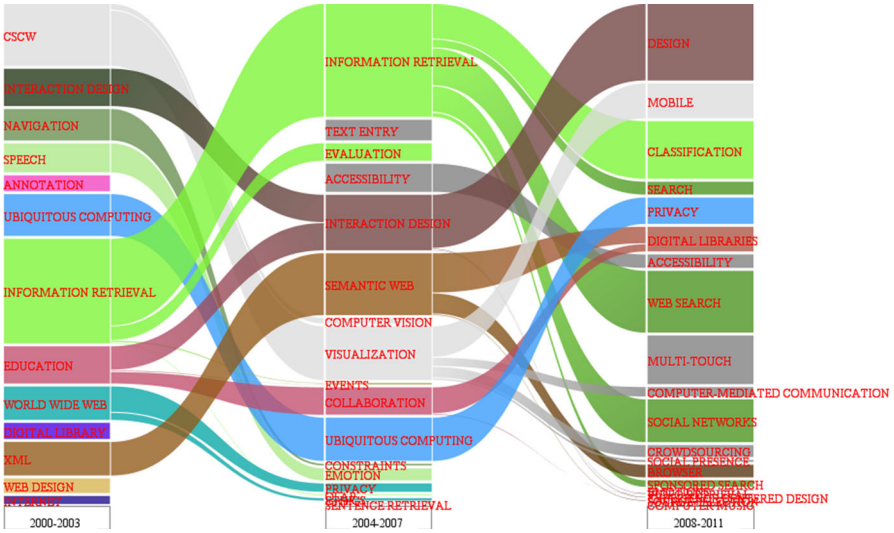

Fig. 4 The global evolution base on FIVE-CONF dataset

global level in a long period, even in a more specific field. We can't see more details, as each dataset has a 3 years span.

In order to discover how each topic evolves, we take the *information retrieval* subject as an example. By focusing the *Information Retrieval* area in T2 in NEViewer, we acquire

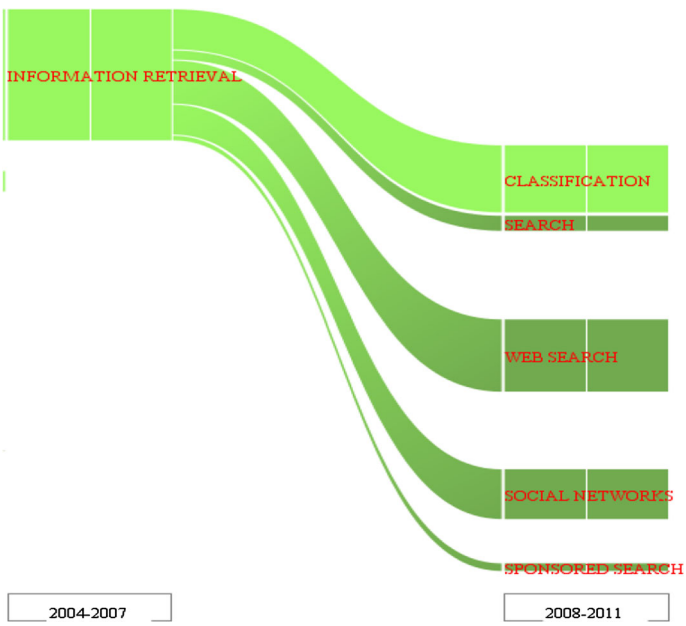

Fig. 5 Evolution of the Topic *Information Retrieval* from T2 to T3

Table 3 Conference sessions of SIGIR from 2008 to 2011

2008	2009	2010	2011
User interaction models	Novel search features	Clustering	Query analysis
Web search	Classification and clustering	User model	Learning to rank
Evaluation	Expansion and feedback	Applications	Retrieval models
Collaborative filtering	Web 2.0	Search engine architectures and scalability	Social media
Learning to rank	Retrieval models	Link analysis & advertising	Web IR
High-performance & high dimensional indexing	Speech and linguistic processing	Learning to Rank	Collaborative filtering
User adaptation & personalization	Recommenders	Filtering and recommendation	Query analysis
Clustering	Question answering	Information retrieval theory	Communities
Multilingual & crosslingual retrieval	Efficiency	Language models & IR theory	Image search
Relevance feedback	Web retrieval	Query representations & reformulations	Web queries
Summarization	Learning to Rank	Automatic Classification	Collaborative filtering
Exploratory search & filtering	Information extraction	Retrieval models and ranking	Multimedia IR
Multimedia retrieval	Click through models	User feedback & User Models	Summarization
Query analysis & models	Vertical search	Web IR and social media search	Query suggestions
Non-topicality	Interactive search	Document structure & adversarial information retrieval	Linguistic analysis
Probabilistic models	Multimedia	Users and interactive IR	Effectiveness
Analysis of social Networks	Federated, distributed search	Document representation and content analysis	Multilingual IR
Question-answering	Industry track speakers	Test-collections	Recommender systems
Social tagging	Evaluation and measurement	Query log analysis	Test collections
Content analysis	Query formulation	Summarization & user feedback	
Learning models for IR	Spamming	Query analysis	
Text classification		Effectiveness measures	
		Multimedia information retrieval	
		Non-English IR & evaluation	

Fig. 5, in which we focuse the process from T2 to T3. In T3, there has five different topics labeled respectively *Classification*, *Search*, *Web Search*, *Social Networks*, and *Sponsored search*.

For evaluating the validity of this process, we manually collected the program sessions of SIGIR conference from 2008 to 2011. The results are shown in Table 3. We find that

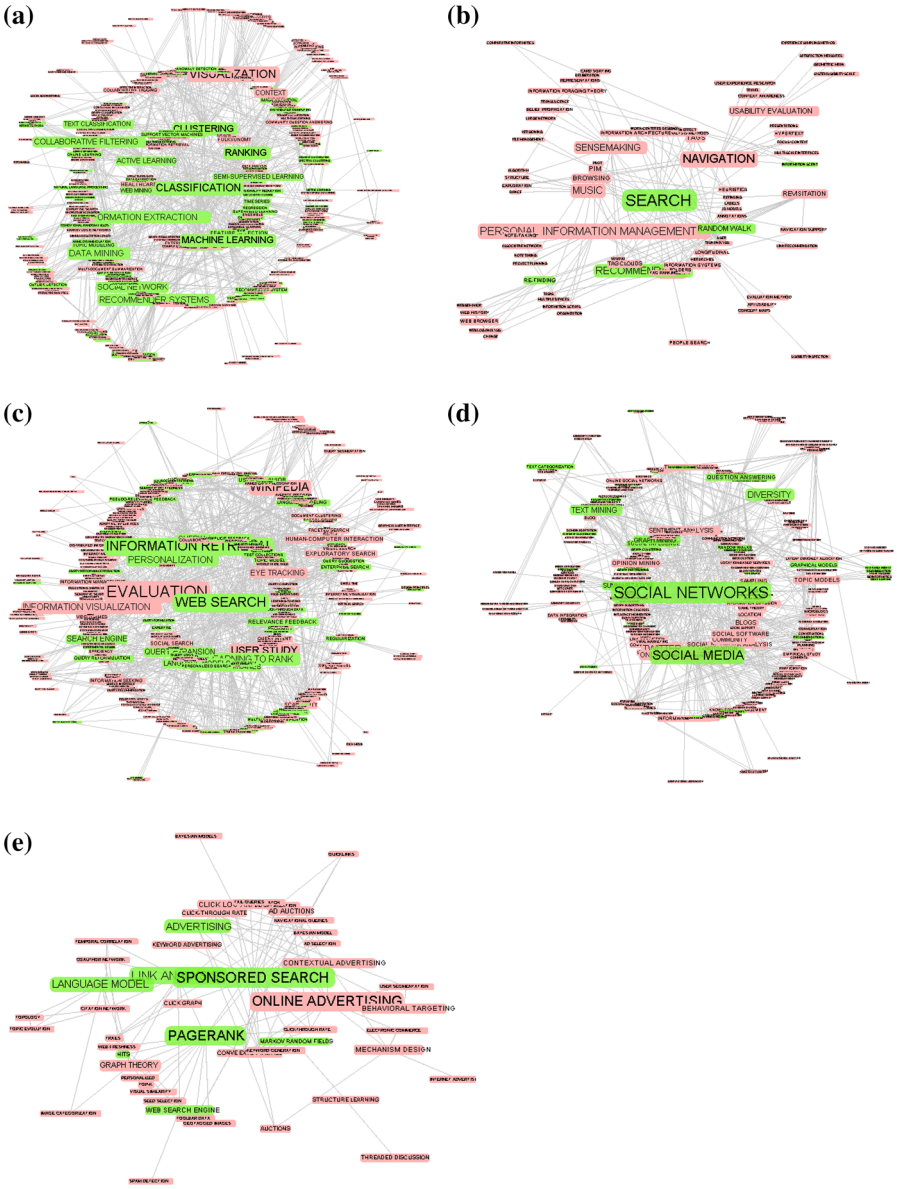

Fig. 6 Backward coloring networks for the five communities. **a** Backward coloring network for *Classification*. **b** Backward coloring network for *Search*. **c** Backward coloring network for *Web Search*. **d** Backward coloring network for *Social Network*. **e** Backward coloring network for *Sponsored Search*

during 2008 and 2010, there is at least one session on *classification*, which can be used to explain the community labeled *classification*. Similarly, almost every year, there is a session named *Web IR* or *Web Search*, which might correspond to the community labeled *Web search*. Social media and Web 2.0 became important in recent years, and almost every

year, there is a session related to *Social media and communities*. We also find a small community named *sponsored search*, this is a comparatively new topic related to the sessions *Link Analysis & Advertising* and *Web IR*. For all of these detected communities, we actually find the similar conference sessions in the real conferences programs, which demonstrate, in a manner, the validity and effectiveness of our work.

In order to respectively reveal the nodes succeeded from the topic *Information Retrieval* in the five communities, five coloring networks are shown in Fig. 6. The sizes of nodes reflect the accumulative frequency that occurred during 2008–2011. In Fig. 6a, we find the community labeled *Classification* consists of many green keywords, such as *machine learning, information extraction, data mining, clustering* and *ranking* etc., which succeeded from the previous topic *Information Retrieval*. Other figures in Fig. 6 look similar. In these diagrams, pink nodes did not succeed from a previous topic *Information Retrieval*. From Fig. 6 we can find that though one node is selected as the label of each community, considering more distinct core nodes (keywords) will help understanding of the corresponding topic of each co-word community.

Discussion

What is the first and vital problem facing to the emerging trend detection and topic evolution analysis? It is how to define a topic and its evolutionary relationship. In the field of information retrieval and data mining, a topic is usually represented by a word cluster including several frequent co-occurring words, named the *bag-of-words* model. In our research framework, the *bag-of-words* model is transferred and adopted, we use the word communities in a co-word network to emblematize research topics. So the topic detection is replaced by community finding in our framework. One or more core nodes in a community, are moreover, skillfully selected as the representative of a corresponding topic. We do not consciously evade the overlapping problems in the community finding phase. In fact, we provide a overlapping algorithm in NEViewer, but we think that it will be more complicated at the later phases when considering the overlapping problem as well, especially at the coloring network visualization phase.

Either from the perspective of informetrics or complex networks, it is a challenge to judge whether there are evolutionary relationships among research topics and what kind of evolutionary forms they take. Firstly, in informetrics, we need a whole comparative analysis including research background, research goals, research methods, many scholars are focused on these two topics; secondly, from the perspective of complex network, we need to carefully analyze nodes, edges, structures even motifs in the two communities.

To simplify the problem, we regard similarity measurement as the most viable way for verifying the evolutionary relationships and offered sevrial similarity indexes in NEViewer. Considering the representing forms of topics, the node-based similarity index is chosen as more comprehensible and adaptable than others to achieve our purposes. A problem facing a relation-based index is the sparsities of real communities. In fact, if the weights of edges are taken into account, the relation-based index will face more conundrums.

Difficulty in choosing a matchable community detection algorithm and preset a appropriate similarity threshold are also challenging problems. A threshold directly determines the verification of community evolution. If we adjust the community detection algorithm and similarity threshold in NEViewer, the alluvial diagram changes accordingly. Therefore, we need to determine how to choose a matchable community detection

algorithm and preset an assortive threshold values for different research fields (science, social science, and humanities) and on multiple research scales (discipline level, subject level, selected topics level). The default threshold value is 0.2 in NEViewer, which can be adjusted by users at any time.

The experimental study conducted with NEViewer has demonstrated six forms of evolution: *Birth*, *Growth*, *Merging*, *Contraction*, *Splitting* and *Death*. In the six forms, *Splitting* and *Merging* all lead to the emergence of new topics. Usually there are some precursors before a new topic emerging in science. To explore the precursors and the causes will help predicting the dynamics of topics and research frontiers. *Growth* means the expansion of research objects under a topic, while *Contraction* reflects the decadence of research topics. *Death* attributes to the constant *Contraction*. These evolutionary states are of great significance in ETD.

At present, several research software tools have been developed for visualizing knowledge structures, research frontiers, hot topics, and research evolution, such as Bibexcel, Science of Science Tool, Citespace, VOSViewer, Network Workbench, SciMat etc. Cobo et al. (2012) analysed the characteristics of these tools in detail. Compared with the existing software tools, NEViewer focuses more on the dynamics of a research field. Through alluvial diagrams and coloring network, the macro evolution processes and micro evolution details are all considered and disclosed in a way.

Conclusion

In this paper, we propose a new research approach based on dynamic co-word networks. In our research framework, we focus on the relationships among keywords other than their frequencies. Community theory is adopted in order to discover the evolution of communities at the mesoscopic level. A powerful software named NEViewer was developed under the guidance of our research approach, which incorporates methods, algorithms, and measures in science mapping workflows. A case study was implemented with the support of NEViewer and the results indicate the existence of six forms of topic evolution.

When compared to the existing research, our work is innovative in three aspects: (a) the design of a longitudinal framework based on the dynamics of co-word communities; (b) it proposes a community labelling algorithm and community evolution verification algorithms; (c) and visualizes the evolution of topics at the macro and micro level respectively using alluvial diagrams and coloring networks. NEViewer can not only be applied to co-word networks, but also to other bibliometric networks, such as cocitation networks, coauthor networks, and even social networks.

As we focus more on network evolution analysis, our research is lacking on the processing of raw data and the building of networks, thus there still remains some disadvantages in our study. First, the algorithm needs to be improved in the phase of constructing co-word networks, second the community evolution verification algorithm is rough to some extent, third the coloring network can not match the overlapping community detection algorithm. However it is an innovation to represent research topics and track their evolution by the dynamics of co-word communities in scientometrics. In the future, we will try more complex and powerful community evolution verification algorithms (Lin et al. 2008; Bródka et al. 2013) to find appropriate similarity thresholds in diverse research fields, choose matched algorithms for varied disciplines, and expand the functions of NEViewer by tracking key events in a specific research field.

Acknowledgments We thank all who helped to improve NEViewer by giving us very valuable suggestions and comments. This project is supported by the National Natural Science Foundation of China (Grant No. 71003078, Grant No. 71173249), the Fundamental Research Funds for the Central Universities, and the Program for New Century Excellent Talents in University.

References

Amitay, E., Carmel, D., Herscovici, M., et al. (2004). Trend detection through temporal link analysis. *Journal of the American Society for Information Science and Technology, 55*(14), 1270–1281.

Ball, B., Karrer, B., & Newman, M. (2011). Efficient and principled method for detecting communities in networks. *Physical Review E, 84*(3), 36103.

Berger-Wolf, T. Y., & Saia, J. (2006). A framework for analysis of dynamic social networks. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, Philadelphia, PA, USA, pp. 523–528.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008.

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37*(1), 179–255.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics, 64*(3), 351–374.

Bródka, P., Saganowski, S., & Kazienko, P. (2013). GED: The method for group evolution discovery in social networks. *Social Network Analysis and Mining, 3*(1), 1–14.

Buente, W., & Robbin, A. (2008). Trends in Internet information behavior, 2000–2004. *Journal of the American Society for Information Science and Technology, 59*(11), 1743–1760.

Chavalarias, D., & Cointet, J. P. (2013). Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PLoS ONE, 8*(2), e54847.

Chen, C. (2005). CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359–377.

Chen, P., & Redner, S. (2010). Community structure of the physical review citation network. *Journal of Informetrics, 4*(3), 278–290.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011a). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology, 62*(7), 1382–1402.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011b). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. *Journal of Informetrics, 5*(1), 146–166.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology, 63*(8), 1609–1630.

Costa, L. F., Rodrigues, F. A., Travieso, G., et al. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics, 56*(1), 167–242.

Ding, Y. (2011). Community detection: Topological vs. topical. *Journal of Informetrics, 5*(4), 498–514.

Garfield, E. (1994). Scientography: Mapping the tracks of science. *Current Contents: Social & Behavioural Sciences, 7*, 5–10.

Goth, G. (2012). The science of better science. *Communication of ACM, 55*(2), 13–15.

Guimerà, R., Sales-Pardo, M., & Amaral, L. A. N. (2006). Classes of complex networks defined by role-to-role connectivity profiles. *Nature Physics, 3*(1), 63–69.

Herrera, M., Roberts, D. C., & Gulbahce, N. (2010). Mapping the evolution of scientific fields. *PLoS ONE, 5*(5), e10355.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp 50–57). New York: ACM.

Klavans, R., & Boyack, K. W. (2006). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology, 57*(2), 251–263.

Kontostathis, A., De, I., Holzman, L. E., & Pottenger, W. M. (2004). Use of term clusters for emerging trend detection. Technical Report. 2004, Available from webpages.ursinus.edu/akontostathis/kontostathisETD.ps.

Kontostathis, A., Galitsky, L., Pottenger, W. M., Roy, S., & Phelps, D. J. (2003). A survey of emerging trend detection in textual data mining. In M. Berry (Ed.), *A comprehensive survey of text mining* (pp. 185–224). Heidelberg: Springer.

Lambiotte, R., & Panzarasa, P. (2009). Communities, knowledge creation, and information diffusion. *Journal of Informetrics, 3*(3), 180–190.

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E, 80*(5), 56117.

Le, M. H., Ho, T. B., & Nakamori, Y. (2005). Detecting emerging trends from scientific corpora. *International Journal of Knowledge and Systems Sciences, 2*(2), 53–59.

Lee, J. Y., Kim, H., & Kim, P. J. (2010). Domain analysis with text mining: Analysis of digital library research trends using profiling methods. *Journal of Information Science, 36*(2), 144–161.

Leydesdorff, L., & Rafols, I. (2008). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology, 60*(2), 348–362.

Lin, Y. R., Chi, Y., Zhu, S., Sundaram, H., & Tseng, B. L. (2008). Facetnet: A framework for analyzing communities and their evolutions in dynamic networks. In *Proceeding of the 17th international conference on World Wide Web*, April 21–25, Beijing, China.

Liu, X., Jiang, T. T., & Ma, F. C. (2013). Collective dynamics in knowledge networks: Emerging trends analysis. *Journal of Informetrics, 7*(2), 425–438.

Mane, K. K., & Börner, K. (2004). Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 5287–5290.

McCain, K. W. (2008). Assessing an author's influence using time series historiographic mapping: The oeuvre of Conrad Hal Waddington (1905–1975). *Journal of the American Society for Information Science and Technology, 59*(4), 510–525.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*(2), 26113.

Palla, G., Barabasi, A. L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature, 446*(7136), 664–667.

Palla, G., Derényi, I., Farkas, I., et al. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature, 435*(7043), 814–818.

Pottenger, W. M., & Yang, T. H. (2001). *Detecting emerging concepts in textual data mining*. In Computational information retrieval. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Price, D. J., & de Solla. (1963). *Little science, big science*. New York: Columbia University Press.

Ronhovde, P., & Nussinov, Z. (2009). Multiresolution community detection for megascale networks by information-based replica correlations. *Physical Review E, 80*(1), 016109.

Rosvall, M., & Bergstrom, C. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proceeding of the National Academy of Sciences of the United States of America, 104*(18), 7327–7331.

Rosvall, M., & Bergstrom, C. (2008). Maps of random walks on complex networks reveal community structure. *Proceeding of the National Academy of Sciences of the United States of America*, *105*(4), 1118–1123.

Rosvall, M., & Bergstrom, C. T. (2010). Mapping change in large networks. *PLoS ONE, 5*(1), e8694.

Roy, S., Gevry, D., & Pottenger, W. M. (2002). Methodologies for trend detection in textual data mining. Proceedings of the Textmine'02 Workshop at the 2nd SIAM Conference on Data Mining. From http://www.cse.lehigh.edu/~billp/pubs/ETDMethodologies.pdf.

Schiebel, E., Hörlesberger, M., Roche, I., et al. (2010). An advanced diffusion model to identify emergent research issues: The case of optoelectronic devices. *Scientometrics, 83*(3), 765–781.

Sun, J., Faloutsos, C., Papadimitriou, S, et al. (2007). GraphScope: Parameter-free mining of large time-evolving graphs. In *The Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 687–696). New York: ACM.

Tu, Y. N., & Seng, J. L. (2012). Indices of novelty for emerging topic detection. *Information Processing and Management, 48*(2), 303–325.

van Eck, N. J., & Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology, 60*(8), 1635–1651.

van Eck, N. J., Waltman, L., Noyons, E. C. M., et al. (2010). Automatic term identification for bibliometric mapping. *Scientometrics, 82*(3), 581–596.

Wallace, M. L., Gingras, Y., & Duhon, R. (2009). A new approach for detecting scientific specialties from raw cocitation networks. *Journal of the American Society for Information Science and Technology, 60*(2), 240–246.

Wang, X. G. (2013). Forming mechanisms and structures of a knowledge transfer network: Theoretical and simulation research. *Journal of Knowledge Management, 17*(2), 278–289.

Wang, X. G., Jiang, T. T., & Li, X. Y. (2010). Structures and dynamics of scientific knowledge networks: An empirical analysis based on a co-word network. *Chinese Journal of Library and Information Science, 3*, 19–36.

Wu, B., Wang, B., & Yang, S. Q. (2011). Framework for tracking the event-based evolution in social networks. *Journal of Software*, *22*(7), 1488–1502.