Scientometrics (2010) 85:689–703
DOI 10.1007/s11192-010-0267-8

Assessment of ontology-based knowledge network formation by Vector-Space Model

Pei-Chun Lee · Hsin-Ning Su · Te-Yi Chan

Received: 5 January 2010 / Published online: 9 July 2010

Abstract This study proposes an empirical way for determining probability of network tie formation between network actors. In social network analysis, it is a usual problem that information for determining whether or not a network tie should be formed is missing for some network actors, and thus network can only be partially constructed due to unavailability of information. This methodology proposed in this study is based on network actors' similarities calculations by Vector-Space Model to calculate how possible network ties can be formed. Also, a threshold value of similarity for deciding whether or not a network tie should be generated is suggested in this study. Four ontology-based knowledge networks, with journal paper or research project as network actors, constructed previously are selected as the targets of this empirical study: (1) Technology Foresight Paper Network: 181 papers and 547 keywords, (2) Regional Innovation System Paper Network: 431 papers and 1165 keywords, (3) Global Sci-Tech Policy Paper Network: 548 papers and 1705 keywords, (4) Taiwan's Sci-Tech Policy Project Network: 143 research projects and 213 keywords. The four empirical investigations allow a cut-off threshold value calculated by Vector-Space Model to be suggested for deciding the formation of network ties when network linkage information is unavailable.

Keywords Social network · Knowledge network · Keyword · Cut-off value · Network formation · Vector-Space Model

P.-C. Lee
SPRU—Science and Technology Policy Research, The Freeman Centre,
University of Sussex, Brighton BN1 9QE, UK

P.-C. Lee · H.-N. Su (✉) · T.-Y. Chan
Science and Technology Policy Research and Information Center, National Applied Research Laboratories, 14 F., No. 106, Sec. 2, He-Ping E. Rd., Taipei 106, Taiwan
e-mail: hnsu@mail.stpi.org.tw; ningsustpi@gmail.com

P.-C. Lee
Graduate Institute of Technology and Innovation Management, National Chengchi University,
No. 64, Sec. 2, Chih-Nan Rd., Wenshan, Taipei 116, Taiwan

Introduction

Mapping ontology-based knowledge network

The basic components of a social network can be different forms of social actors, for example, individuals, organizations, countries. A social network formed on the basis of social exchange can be used for understanding how resources are exchanged, how social actors are positioned to influence resource exchange, and which resource exchange is important (Nohria et al. 1992; Wasserman and Galaskiewicz 1994; Wellman and Berkowitz 1988). Each of resource exchange is a social network or a "tie" maintained by social actors at both end of the "tie", the strength of a tie as a function of the number of resource exchange, the type of exchange, the frequency of resource exchange, or even how close the two connected actors are (Marsden and Campbell 1984).

Social network analysis is an interdisciplinary research field. Granovetter proposed the theory of weak tie after his survey of 282 workers in regards to the type of ties between the job changer and the contact person who provided the necessary information. Of those who found jobs through personal contacts, only 16.7% reported seeing their contact often (Granovetter 1970, 1973). This illustrates social network analysis is a proxy which provides interconnection between microscopic analysis and macroscopic analysis. In the late 1990s, collaboration between researchers from different fields by the use of social network analysis had been initiated so social network analysis become more interdisciplinary. Barabasi and Albert demonstrated that the algebraic distribution in the connectivity of scale-free network is caused by two basic factors in the temporal evolution of the network: growth and preferential attachment (Barabasi and Albert 1999). Watts and Strogatz contributed to expansion of small world concept from conventional neuro-science and bio-information system to any natural or human system that can be modeled by network (Watts and Strogatz 1998; Watts 2003).

Motter et al. constructed a conceptual network from the entries in a thesaurus dictionary considering two words connected if they express similar concepts. He argued that language network exhibits the small-world property as a result of natural optimization and these findings are important not only for linguistics, but also for cognitive science (Motter et al. 2002). Marshakova-Shaikevich built a semantic map of a field of women's studies by document clustering on the basis of lexical similarity of titles and word clustering on the basis of co-occurrence of words in the same documents (Marshakova-Shaikevich 2005). Similar investigations have been done to map different knowledge fields by keyword-based method, for example, Biochemistry (Rip and Courtial 1984), Chemistry (Callon et al. 1991), Neural network research (Van Raan and Tijssen 1993; Noyons and Van Raan 1998), Biological Safety (Cambrosio et al. 1993), Optomechatronics (Noyons and van Raan 1994), Adverse drug reactions (Clarke et al. 2007), Software engineering (Coulter et al. 1998), etc.

To improve knowledge mapping where conceptualization of knowledge is essential, the concept of ontology can be applied to knowledge mapping. Neches defined ontology as basic terms and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary (Neches et al. 1991). The general basis for ontology is the emphasis on shared vocabularies and on properties that hold in all situations. Ontologies are content theories about the sorts of objects, properties of objects, and possible relations between objects in a specified domain of knowledge (Chandrasekaran et al. 1999). Thus, informally defined, "ontologies are agreements about shared conceptualizations".

The use of ontology provides a conceptual description for knowledge domain. Its primary function is to provide humans with a framework for interacting with application systems, through the use of ontology to improve the communication model between humans and machines (Weng and Chang 2008). Considerable ontology researches have been recently conducted to allow computer to provide information with knowledge implication. For example, Liao et al. (2005) proposed a framework in which semantic web ontology is used to model the contexts, user profiles, and product/service information (Liao et al. 2005), Pirró and Talia (2010) proposed an ontology mapping system with strategy prediction capabilities, Zhang et al. (2009) proposed an ontology and peer-to-peer based data and service unified discovery system (Zhang et al. 2009). Weng and Chang (2008) utilized ontology network analysis for research document recommendation (Weng and Chang 2008). This study reaps rewards from ontology and evaluations knowledge network based on the concept of ontology. For an ontology-based knowledge network in this study, vocabularies of a topic area are author keywords of papers or projects in a specific area, and relations between objects are network linkages based on cooccurrence of same keywords comprised in different papers or projects. The ontology-based knowledge networks studied in this research not only allow users to understand the knowledge structure of a selected domain, but also provide diverse applications e.g. identification of partnership, identification of competitor, document recommendation, reviewer recommendation, etc.

How to determine formation of an ontology-based knowledge network?

The formation of network relies critically on understanding whether or not any two actors in the network should be linked, the strength of linkage between the two actor is proportional to the degree of resource exchange between the two actors. However, in a knowledge network with research documents as network actors, whether or not any two research documents should be linked relies on the similarity of the two documents. The similarity between any two documents falls between 0 and 1 (0–100%), but in order to facilitate quantitative network centralities calculation (Granovetter 1973; Freeman 1979; Brass and Burkhardt 1992), the similarity values between 0 and 1 have to be transformed into 0 or 1 to represent networking behaviors ("0" refers to no linkage between two actor, "1" refers to formation of linkage between two actors). By doing so, diverse applications based on quantitative network centralities calculation are then possible, e.g. network visualization, knowledge map, performance evaluation, resource allocation, etc. In this sense, a cut-off value has to be obtained, so similarity values smaller than the cut-off value will be converted to "0" and for those similarity values larger than the cut-off value will be converted to "1".

There are many types of bibliometric information based on which knowledge network can be created, e.g. co-keyword, co-author, citation, etc. However, co-keyword method which we believe is a critical approach, since keyword is the most basic fundamental carrier of knowledge, is selected in this study for evaluating ontology-based knowledge network. The three major steps for determining formation of an ontology-based knowledge network comprise:

(1) *Create knowledge network by co-keyword based method*: Relation between two different documents occurred because these two documents contain at least one same author keyword. This is based on an assumption made in knowledge network that keyword is the basic knowledge carrier and any two documents sharing same

keyword implies they are partially overlapped in a knowledge area that can be represented by that keyword. (Keywords used are author keywords.)

(2) *Document similarity calculation*: Document similarity is obtained by calculating the similarity of document abstracts by the Vector-Space Model (Salton et al. 1975; Salton and McGill 1983; Raghavan and Wong 1986) which is an application model for information filtering, information retrieval, document indexing, and coefficient evaluation. The VSM is used to analyze document abstract by choosing author keywords as the set of words which appeared in the document. This provides two advantages: (1) abstract is the core of a document, the utilization of abstract instead of full-text provides a fair basis for keyword frequency calculation in VSM, because this avoids the problem on keyword frequency calculation caused by document size differences, (2) the utilization of author keywords which serve as representatives of core concepts of knowledge can avoid the consideration of stop words when calculating document similarities.

(3) *Cut-off value determination*: The knowledge network created by co-keyword method, acted as the standard network, is served as the basis to be compared with networks created by choosing different cut-off values. Document similarity values which are larger than the cut-off value will be converted to "1" indicating the two documents are linked together. The best cut-off value which provides maximum match between knowledge network created by co-keyword method and knowledge network created by choosing different cut-off values is suggested as the optimized value for representing overall networking behaviors.

Research question

The formation of an ontology-based knowledge network relies critically on the selection of cut-off value. Different cut-off value leads to different network structure and network properties. By increasing the cut-off value, network ties which similarity value is not higher than the cut-off value will be removed. The question raised here is what is the best cut-off value? How do we optimize cut-off value?

Yoon and Park argued that "The determination of cutoff value is in nature a subjective, trial-and-error task. The network becomes denser as the cutoff value becomes lower, whereas it becomes sparser as the cutoff value becomes higher. At some intermediate level, the analyzer has to select a reasonable value so that the structure of the network becomes clearly visible" (Yoon and Park 2004). Cavenago et al. also suggested that "There are no established criteria for selecting a cut-off, we examined different minimum average tie strengths: after evaluating different alternatives, we choose a cut-off value 0.5″ (Cavenago et al. 2009). Schildt and Mattsson (2006) used 0.25, Schildt et al. (2006) used 0.15 for their cut-off value selections in their researches.

It is suggested in literature that cut-off value is an arbitrary value for studies in bibliometrics analysis. However, we should be able to optimize the selection of cut-off value by the aid of useful information available in research samples. For ontology based-knowledge network study, document similarity calculation can be served as evidences for determining how close any two documents are. By the aid of useful information, i.e. author keywords, we will be able to create the basis for optimizing cut-off value associated with knowledge network structure. Subsequently, we integrate author keyword and document similarity calculation to illustrates a possibly way of optimizing formation of an ontology-based knowledge network, also answer the question of how to obtain the best cut-off value in a knowledge network.

Data preparation

The data sources used in this research are four knowledge fields formed by research papers or research projects. Four cases investigated in this study are: (1) Technology Foresight (Su and Lee 2009a), (2) Regional Innovation System, (3) Global Sci-Tech Policy (Lee et al. 2009a), (4) Taiwan Sci-Tech Policy (Su et al. 2009). The rationales why the four knowledge fields are chosen is because:

(a) They are all fields that are in urgent need of collaboration and interdisciplinary exchanges, which fit the purpose of investigating knowledge network creation (e.g. identifying research partnership, identification of research competitor).
(b) Even though cut-off vale determination is regarded trial and error, the methodology proposed in this study for optimizing cut-off value maybe sensitive to research field. It is better to select case studies from the same or similar fields. The four cases all belong to the same field of national level planning.
(c) We are highly interested on national level planning and our background and practical experiences enable us to obtain much deeper insight on knowledge network related to national planning issues. Also, our practical experiences on national level planning allow us to test the validity of the cut-off value optimization proposed in this study.

Four knowledge networks representing the four knowledge fields are constructed, in order to serve as bases to be compared by simulated network subsequently. The constructed networks and their corresponded adjacency matrices are named as "standard network" and "standard adjacency matrix" in this study. Figure 1 shows the process of constructing standard.

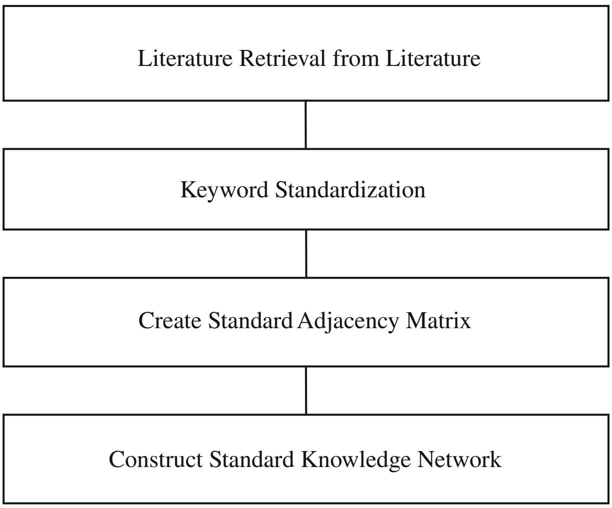

Fig. 1 Processes of creating standard knowledge network

Literature retrieval

Knowledge networks (1)–(3) are based on research papers for the fields of Technology Foresight, Regional Innovation System and Global Sci-Tech Policy, respectively. Research papers are retrieved from online literature databases, e.g. Web of Science, Emrald, Inderscience, by the use of different query strategies. Knowledge network (4) is based on research projects supported by Taiwan government and are retrieved from Government Research Bulletin Database (National Science Council 2009).

Keyword standardization

For the three knowledge networks, Technology Foresight, Regional Innovation System, and Global Sci-Tech Policy, keywords are author keywords assigned by authors in each research paper. For the community of Taiwan Sci-Tech Policy, the author keywords are not well assigned so title keywords are used. Title of each research project report is parsed to have keywords retrieved for subsequent calculation.

Due to the fact that different words can be used for describing the same meaning, it is necessary to standardize words that used to express the same meaning. For example, (1) plurality form of word is standardized to its singularity form; (2) technique, technologies, technology are standardized to technology; and (3) regional systems of innovation, RIS, and industrial cluster are standardized to "regional innovation system". The top 15 highest occurrence standardized keywords for the four communities are listed in Table 1.

Creation of standard adjacency matrix

The method of establishing network in this study is based on author or title keywords and keyword co-occurrence, and defined as research focus parallelship network (RFP network). For an RFP network created in this study, relation between two different papers occurred

Table 1 Top 15, keywords in the four knowledge networks

Ranking	Technology Foresight	Regional Innovation System	Global Sci-Tech Policy	Taiwan Sci-Tech Policy[a]
1	Foresight	Innovation	Sci-Tech policy	Sci-Tech policy
2	Technology Foresight	Regional development	Policy	Sci-Tech policy research
3	Technology	Regional innovation system	Innovation	Planning
4	Forecasting	Cluster	R and D	Supervision and evaluation
5	Strategic planning	Regions	Climate change	Sci-Tech program
6	Delphi Method	R&D	Technology	Industry
7	Innovation	Network	Science policy interface	Government
8	Scenario planning	Biotechnology	Science	Science and Technology
9	Innovation policy	Entrepreneurship	Environmental policy	Innovation
10	Research	Innovation system	Technology transfer	Strategy

[a] Keywords of Taiwan Sci-Tech Policy are translated from Chinese

because these two documents share at least one same keyword. For example, document is used as a network actor (network node) and any of two actors sharing one same keyword will be connected. This is based on an assumption made in this study that keyword represents the core of research of a document and any two documents sharing same keyword implies these two researches are partially overlapped in an area that can be represented by that keyword. The two documents are thus regarded as a pair of parallel documents and the constructed network is defined as RFP network (Su and Lee 2009a, b). An adjacency matrix can therefore be created by the above co-keyword based analysis on document linkages., entries in the created adjacency matrix are assigned as either 1 or 0 to represent two paired documents are connected or disconnected, respectively.

Constructing standard network

After creation of standard adjacency matrix, a standard network can be constructed by computer software to obtain visualization of this four research communities. Figure 2 illustrates four knowledge networks investigated in this study. Table 2 shows basic network information for the four knowledge networks.

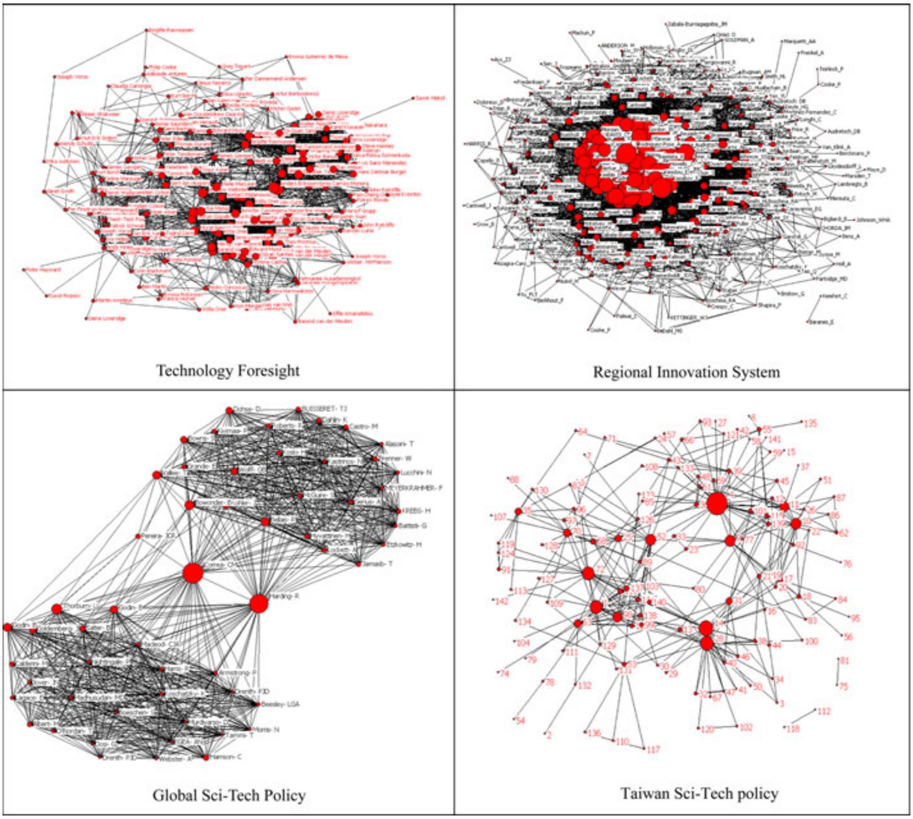

Fig. 2 Four standard networks analyzed in this study

Table 2 Information of four standard networks

	Technology Foresight	Regional Innovation System	Global Sci-Tech Policy	Taiwan Sci-Tech Policy
No. of paper/project	181	431	548	143
No. of keywords (with duplication)	883	2053	3597	598
No. of keywords (without duplication)	554	1165	2377	213
No. of ties	2399	9031	7740	328

Method

Document similarity calculated by VSM

There are ways to calculate the similarity between documents, such as Dice coefficient, Jaccard coefficient and Cosine similarity (van Rijsbergen 1979). The first two measurements involve both the keywords intersection and union and the Cosine similarity calculates the angle between two term vectors (documents).

Since this study transformed documents into term vectors in defined dimensions (i.e. the chosen 546 keywords), this study adopts the famous Cosine similarity to measure the similarity between documents. As mentioned, the Cosine similarity is used to measure the similarity of two vectors of certain dimension space via finding the cosine angle of two vectors, and it is often used to compare the documents in VSM. Formula (3) shows the similarity calculation between two vectors A and B of n dimensions (i.e. n terms selected in steps of feature selection).

$$A = (W_{A1}, W_{A2}, W_{A3}, W_{A4}, \ldots, W_{An}) \tag{1}$$

$$B = (W_{B1}, W_{B2}, W_{B3}, W_{B4}, \ldots, W_{Bn}) \tag{2}$$

$$\text{Similarity}(A, B) = \cos(A, B) = \frac{\sum_{i=1}^{n} (W_{Ai} \cdot W_{Bi})}{\sqrt{\sum_{i=1}^{n} (W_{Ai})^2} \times \sqrt{\sum_{i=1}^{n} (W_{Bi})^2}} \tag{3}$$

Since the similarity is a cosine value, the calculated similarity value is between 0 and 1, the larger the similarity value the more similar the documents are. By the use of this method, a similarity matrix of scores which expresses the similarity between any two documents can be obtained.

Convert similarity matrix into adjacency matrix

In order to construct knowledge network, a binary adjacency matrix is required. The binary value of entry (i, j) in the matrix equals to 1 if the link between node i and node j is considered strong, and equals to 0 if the link between node i and node j is considered week. Whether the linkage between any two nodes is 0 or 1, is decided by a cut-off value that is to be determined in this study, i.e. any two nodes are linked if the similarity obtained by Formula (3) is higher than the cut-off value and the binary value is set as 1 in the adjacency matrix. Otherwise, the two nodes are considered not linked and the binary value is set as 0 in the adjacency matrix.

Figure 3 is an example illustrating the process of converting from a similarity matrix to a network structure by selecting a cut-off value of 0.4. All elements in the similarity matrix smaller than 0.4 are converted to 0 in the adjacency matrix, the other elements are larger than 0.4 and therefore converted to 1 in the adjacency matrix. Based on the adjacency matrix, a network structure can be constructed by conventional social network analysis method.

Characteristics of cut-off value

As shown in Fig. 3, once if the cut-off value can be decided, the similarity matrix can be converted to adjacency matrix very easily. The question here is how to decide the cut-off value. Two important characteristics when optimizing a cut-off value is: (1) The cut-off value should not be a universal value that can be applied on any conversion of similarity matrix to adjacency matrix, therefore different cut-off values should be assigned in different cases, (2) ideally there should be multiple cut-off values in one single conversion but this will lead to complication of matrix conversion. After considering the above two characteristics, it is more desirable to assume one matrix conversion only takes one cut-off value, and different cut-off value should be assigned for different matrix conversion. In the sense of the purpose of mapping knowledge networks in this study, one cut-off value assigned for one knowledge networks of same research field and same document type is desirable, otherwise research connectivity or network tie assessment related researches, i.e. identification of research collaborator, competitors or peer reviewer, would not be possible.

Example

Similarity matrix

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
Doc1	1	0.2	0.6	0.7	0.6	0.1
Doc2	0.2	1	0.3	0.8	0.5	0.6
Doc3	0.6	0.3	1	0.2	0.3	0.2
Doc4	0.7	0.8	0.2	1	0.9	0.7
Doc5	0.6	0.5	0.3	0.9	1	0.8
Doc6	0.1	0.6	0.2	0.7	0.8	1

Cut-off Value = 0.4 →

Adjacency matrix

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
Doc1	0	0	1	1	1	0
Doc2	0	0	0	1	1	1
Doc3	1	0	0	0	0	0
Doc4	1	1	0	0	1	1
Doc5	1	1	0	1	0	1
Doc6	0	1	0	1	1	0

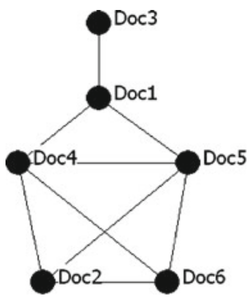

Fig. 3 The process of matrix conversion for creating network structure

Conventional way of determining cut-off value is an arbitrary and trial-and-error task. The density or the number of network ties of a network can be very high if a low cut-off value is set, the density or the number of network ties of a network can be very low if a high cut-off value is set. It is always a question that needs to be answered by social network researcher—What is the best cut off value for an investigated network. Usually it is suggested to apply multiple values for a sensitivity analysis, and then an arbitrary cut-off value can be selected by considering whether the structure of a network can be reasonably visualized (Yoon and Park 2004; Lee et al. 2009b; Yoon et al. 2008). However, to avoid the above conventional way of determining cut-off value in arbitrary and subjective manners, this study proposes a method which allows a more objective determination of cut-off value for social network research.

Optimization of cut-off value

This study transforms the calculated document similarity values into connectivity in different systems. Even though the similarity value is a cosine value, formula (3) should be distributed between 0 and 1, the minimum similarity value is not necessary to be 0 and the maximum similarity value is not necessarily to be 1, either. In order to allow objective comparison among different research communities with different distribution of similarity values, this study normalizes the distribution of similarity values to allow the minimum similarity to be 0 and maximum similarity value to be 1. The selected normalized cut-off value is thus between 0 and 1 so cut-off value comparison cross different research communities is possible.

By choosing different normalized cut-off values from 0 to 1 with accumulation of interval 0.01, 100 different simulated networks for each research community are constructed. The way of constructing network by VSM similarity calculation and varying cut-off value is illustrated in Fig. 3. The constructed 100 networks for each knowledge network are called "pseudo-networks" and their corresponded adjacency matrices are named as "pseudo adjacency matrices". The 100 pseudo adjacency matrices are compared with the standard adjacency matrix by calculating the similarity of each entry between the standard adjacency matrix and the 100 pseudo adjacency matrices.

It is believed that the connection and disconnection between nodes in a social network carry equal important information, i.e. the importance of reason why two nodes are connected or disconnected are the same. Therefore, a simple match coefficient (Sokal and Michener 1958) which can give consideration to both connected and disconnected nodes of reference social network is adopted to calculate the similarity between standard adjacency matrix and pseudo adjacency matrix (or between standard network and pseudo networks). Formula (4) shows the simple match coefficient used in the matrix similarity calculation.

$$S_{ij} = \frac{a + d}{a + b + c + d} \tag{4}$$

a is the number of the entries that are 1 in both standard adjacency matrix and pseudo adjacency matrix, b is number of the entries that are 1 in standard adjacency and 0 in pseudo adjacency matrix, c is number of the entries that are 0 in standard adjacency and 1 in pseudo adjacency matrix, and d is the number of the entries that are 0 in both standard adjacency matrix and pseudo adjacency matrix.

Result

The calculated similarity between standard adjacency matrix and pseudo adjacency matrix is shown as percentage of match to standard adjacency matrix by varying cut-off value. Figures 4, 5, 6, and 7 are percentage of matrix match versus cut-off value for Technology Foresight (Su and Lee 2009a), Regional Innovation System, Global Sci-Tech Policy (Lee et al. 2009a), and Taiwan Sci-Tech Policy (Su et al. 2009), respectively.

Exponential increase can be observed in Figs. 4, 5, and 6, the matches increase as the cut-off values increase for the three cases. However, the increase of cut-off value, when

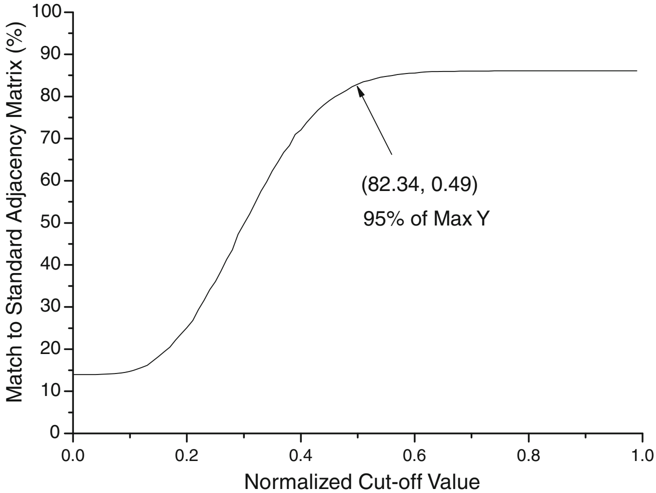

Fig. 4 Matrix match versus cut-off value for Technology Foresight

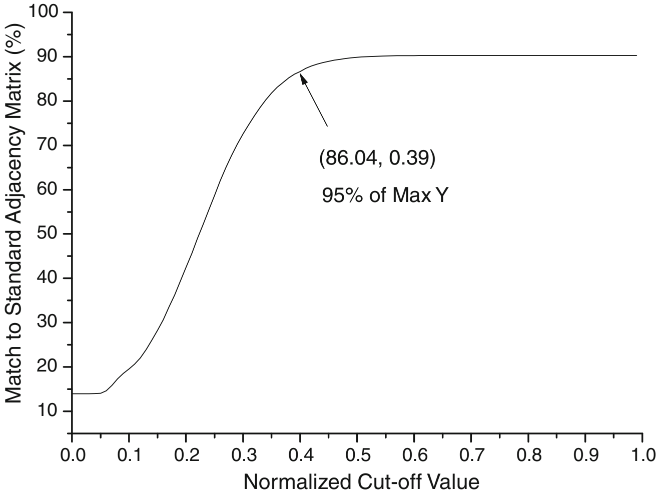

Fig. 5 Matrix match versus cut-off value for Regional Innovation System

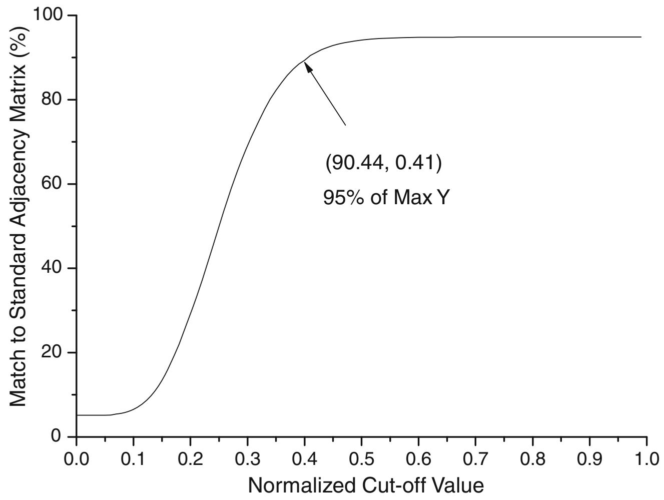

Fig. 6 Matrix match versus cut-off value for Sci-Tech Policy

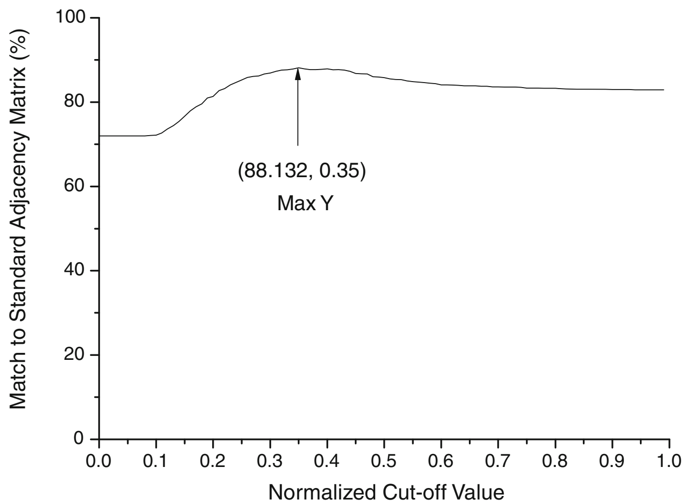

Fig. 7 Matrix match versus cut-off value for Taiwan Sci-Tech Policy

cut-off value is higher than about 0.5, brings very insignificant effect to percentage of match so the curves become as flat as horizontal lines. To avoid this inefficient cut-off value increase which leads to almost no contribution to percentage of match and can possibly ruin the application of cut-off value to other systems, the cut-off values where 95% of maximum matches are selected as the best cut-off values. The obtained best cut-off values are 0.49, 0.39 and 0.41 for the three cases in Figs. 4, 5, and 6.

In Fig. 7, a hump can be observed so it is very straight forward to select the tip of the hump as the place where maximum match and the best cut-off value (0.35) are presented.

Application of cut-off value-identification of potential research partnership

This use of VSM for similarity calculation is based on a set of author keywords or title keywords that represent a research field or a research interest. The set of keywords can be modified by researchers according to their research interest, and then use VSM to calculate similarity of documents collected from data sources, e.g. Web of Science Database, Patent Database, Books. Finally the cut-off value obtained previously can be applied to decide whether or not any two documents should be linked or not. Any paired documents imply potential opportunity to have collaboration between the two authors of the paired documents, because of the high similarity between their documents.

Similarly, this method can also be applied to any situation when identification of strong ties between actors in a networked system is necessary. For example, identification of competitor, selection of peer reviewer or committee member, discover potential research interests, etc.

Discussion and conclusion

The obtained curves in Figs. 4, 5, 6, and 7 for determination of cut-off value are either exponentially or hump-shaped. The exponential curves in Figs. 4, 5, and 6 are for the three cases that utilize paper abstract as data source, but Fig. 7 is the case that utilizes Chinese research project report. We should not conclude that any correlation existed between data source and curve shapes, but different type of data source may possibly contribute to different curve shapes. However, this is very empirical and it requires more studies on different research communities to have any further conclusion.

The determination of cut-off value is an arbitrary and trial-and-error task. The four case studies selected in this research are used to demonstrate a way for determining cut-off values, it is not intended and also impossible to calculate a cut-off values that can be applied universally. It is suggested that researcher should calculate their own cut-off values since the cut-off value is very case dependent. Different data source and different set of keyword representing research interests will lead to different cut-off values. Also, the cut-off value should be a function of time because human knowledge system is of no doubt dynamic.

This study utilizes paper abstract or title of project report as data for VSM calculation, it is possible to use the method proposed in this study to calculated similarities among unstructured data, but the number of words in each record should not be too different. VSM calculation considers the number of times keywords are presented in the document. It would lead to significant bias when comparing very long documents with very short documents. A possible solution is to normalize the calculated similarity values by considering the total number of words appeared in each document.

The cut-off value should be a task-based and case dependent variable. There is no cut-off value that can be applied universally and the best cut-off value is always remained unknown. Since conventional determination of cut-off value is very subjective and arbitrary. This study proposes a method to optimize cut-off value in a more systematic and objective way in order to identify potential connectivity between any two actors in a networked system.

This cut-off value optimization method provided in this study can be directly used to determine knowledge network as well as identify potential collaborator or competitor. However, the main purpose of this study is to have a comprehensive demonstration on optimizing cut-off value with the aid of useful information available on research samples. It if of great importance for social scientists of science to take advantage of additional information to approach better to the best cut-off value, so the selection of cut-off value is no longer trial and error but possibly a little more objective than traditional methods. In addition, the concept of utilizing additional information to have more objective approach is not necessarily for constructing knowledge network, whatever tasks rely heavily on empirical estimation may have potential to obtain more objective approach if some useful information can be considered.

References

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512.

Brass, D. J., & Burkhardt, M. E. (1992). Centrality and power in organizations. In N. Nohria & R. Eccles (Eds.), *Networks and organizations: Structure, form and action* (pp. 191–215). Boston: Harvard Business School Press.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics, 22*(1), 155–205.

Cambrosio, A., Limoges, C., Courtial, J. P., & Laville, F. (1993). Historical scientometrics? Mapping over 70 years of biological safety research with coword analysis. *Scientometrics, 27*(2), 119–143.

Cavenago, D., Marafiot, E., Mariani, L., & Trivellato, B. (2009). *Network governance and evaluation in public services: A bibliometric literature review*. Presented at the conference of the European Group of Public Administration (EGPA), Malta. Retrieved from http://surplus-unibic.cilea.it/oa/handle/10281/8901.

Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent Systems, 14*(1), 20–26.

Clarke, A., Gatineau, M., Thorogood, M., & Wyn-Roberts, N. (2007). Health promotion research literature in Europe 1995–2005. *The European Journal of Public Health, 17*(Supplement 1), 24–28.

Coulter, N., Monarch, I., & Konda, S. (1998). Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science, 49*(13), 1206–1223.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks, 1*(3), 215–239.

Granovetter, M. S. (1970). *Changing jobs: Channels of mobility information in a suburban community*. Unpublished doctoral dissertation, Harvard University, Boston, MA.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology, 78*(6), 1360–1380.

Lee, S. Y., Su, H. N., Lee, P. C., & Chan, T. Y. (2009a). *Mapping global science and technology policy research structure*. Presented at the 2009 annual conference of the Chinese Society for Management of Technology, Taipei, Taiwan.

Lee, S., Yoon, B., Lee, C., & Park, J. (2009b). Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping. *Technological Forecasting & Social Change, 76*(6), 769–786.

Liao, L., Xu, K., & Liao, S. S. (2005). Constructing intelligent and open mobile commerce using a semantic web approach. *Journal of Information Science, 31*(5), 407.

Marsden, P. V., & Campbell, K. E. (1984). Measuring tie strength. *Social Forces, 63*(2), 482–501.

Marshakova-Shaikevich, I. (2005). Bibliometric maps of field of science. *Information Processing and Management, 41*(6), 1534–1547.

Motter, A. E., de Moura, A. P. S., Lai, Y. C., & Dasgupta, P. (2002). Topology of the conceptual network of language. *Physical Review E, 65*, 065102.

National Science Council. (2009). Government research bulletin. Retrieved from http://www.grb.gov.tw.

Neches, R., Fikes, R. E., Finin, T., Gruber, T., Patil, R., et al. (1991). Enabling technology for knowledge sharing. *AI Magazine, 12*(3), 36.

Nohria, N., Eccles, R. G., & School, H. B. (1992). *Networks and organizations: Structure, form, and action*. Boston, MA: Harvard Business School Press.

Noyons, E. C. M., & van Raan, A. F. J. (1994). Bibliometric cartography of scientific and technological developments of an R & D field. *Scientometrics, 30*(1), 157–173.

Noyons, E. C. M., & Van Raan, A. F. J. (1998). Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American Society for Information Science, 49*(1), 68–81.

Pirró, G., & Talia, D. (2010). UFOme: An ontology mapping system with strategy prediction capabilities. *Data & Knowledge Engineering, 69*(5), 444–471.

Raghavan, V., & Wong, S. (1986). A critical analysis of the vector space model for information retrieval. *Journal of the American Society for Information Science and Technology, 37*(5), 279–287.

Rip, A., & Courtial, J. P. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics, 6*(6), 381–400.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval* 1. *Information Processing & Management, 24*(5), 513–523.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw Hill.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 620.

Schildt, H. A., & Mattsson, J. T. (2006). A dense network sub-grouping algorithm for co-citation analysis and its implementation in the software tool Sitkis. *Scientometrics, 67*(1), 143–163.

Schildt, H. A., Zahra, S. A., & Sillanpaa, A. (2006). Scholarly communities in entrepreneurship research: A co-citation analysis. *Entrepreneurship Theory and Practice, 30*(3), 399–415.

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin, 38*, 1409–1438.

Su, H. N., & Lee, P. C. (2009a). *Mapping knowledge evolution of technology foresight*. Presented at the Academy of Management, Chicago, IL, USA.

Su, H. N., & Lee, P. C. (2009b). Knowledge map of publications in research policy. In *Portland international conference on management of engineering & technology*, Portland, OR, USA.

Su, H. N., Lee, P. C., Chien, I. C., & Chan, T. Y. (2009). *Network perspective of science and technology policy research community in Taiwan*. Presented at the 2009 annual conference of the Chinese Society for Management of Technology, Taipei, Taiwan.

Van Raan, A. F. J., & Tijssen, R. J. W. (1993). The neural net of neural network research: An exercise in bibliometric mapping. *Scientometrics, 26*(1), 169–192.

van Rijsbergen, C. J. (1979). *Information retrieval*. London: Butterworths.

Wasserman, S., & Galaskiewicz, J. (1994). *Advances in social network analysis: Research in the social and behavioral sciences*. Newbury Park: Sage.

Watts, D. J. (2003). *Six degrees: The science of a connected age*. New York: WW Norton & Company.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440–442.

Wellman, B., & Berkowitz, S. D. (1988). Introduction: Studying social structures. In B. Wellman & S. D. Berkowitz (Eds.), *Social structures: A network approach* (pp. 1–14). Cambridge: Cambridge University Press.

Weng, S., & Chang, H. (2008). Using ontology network analysis for research document recommendation. *Expert Systems with Applications, 34*(3), 1857–1869. doi:10.1016/j.eswa.2007.02.023.

Yoon, B., Lee, S., & Lee, G. (2008). *Keyword-based knowledge map of academic research*. Presented at the third European conference on management of technology, Nice, France.

Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *Journal of High Technology Management Research, 15*(1), 37–50.

Zhang, Y., Qu, Y., Huang, H., Yang, D., & Zhang, H. (2009). An ontology and peer-to-peer based data and service unified discovery system. *Expert Systems with Applications, 36*(1), 5436–5444. doi: 10.1016/j.eswa.2008.06.083.