Literature-based Multidiscipline Knowledge Discovery: A New Application of Bibliometrics

Jinyan Su[1] and Chunlei Zhou[2]

[1] *sujinyanyc@hotmail.com*
School of Information Management, Wuhan University, Hubei Province (China)

[2] *zclicq@mail.whu.edu.cn (correspondent)*
School of Information Management, Wuhan University, Hubei Province (China)

Abstract

We present a new application of bibliometrics, Literature-based Multidiscipline Knowledge Discovery (LMKD), which is quite different from the well known literature-based knowledge discovery method given by Don R. Swanson. The goal of LMKD is to discover new, potentially meaningful academic visions and relations among given disciplines, by mining bibliographic databases (here we take the case of CSSCI). By using LMKD method, we can try to find the similarities and differences between two research regions and the nature of the disciplines' relationship will be known clearly. As a discovery method, LMKD can be used in a more widely research area than other literature-based discovery methods.

1. Introduction

Information overloading has become a significant problem for researchers in the world. On one hand, scientific literature is readily available, but the sheer volume and growth rate of the literature make it impossible for researchers to keep up with new findings outside their own narrowing fields of expertise. On the other hand, "Verification of results that logically related non-interactive literatures are potential sources of new knowledge (Swanson, 1989a)." Therefore, new methods are needed to help researchers capture and explore new knowledge in the literature.

In this article, we contribute to Literature-based Multidiscipline Knowledge Discovery (LMKD) by proposing a five-step model of discovery in which new scientific hypotheses can be generated and subsequently tested.

2. Background

2.1 Related Research Abroad

During the past two decades, Don R. Swanson has advanced a different view of creating new scientific knowledge. He proposes that combining existing bibliographic information, though not connected, results in new knowledge. One publication may state the relation between the two phenomena A and B, while another reports on the relation between the phenomena B and C. If no one has reported on the association between A and C, this association can be considered to be new and may be of scientific interest. The crucial notion in this view is that two pieces of information are not related directly: there is only a hidden connection. One or more common aspects of the two pieces provide indirect linking. Once these links have been found, a connection can be made and new knowledge be created.

Since 1986, Swanson has made literature-based discovery (LBD) on a regular basis in the scientific field of biomedicine. The first discoveries (Swanson, 1986, 1987, 1988, 1989b) have been corroborated experimentally and clinically (Smalheiser & Swanson, 1998). Gordon and Lindsay (Gordon & Lindsay, 1996) and Weeber et al. (Weeber etc., 2001) have repeated some of Swanson's discoveries with different methods. Weeber et al. have also discovered several hypothetical new therapeutic applications of existing drugs (Weeber etc., 2003). Recently, Srinivasan also replicated most of the Swanson and Smalheiser's discoveries with a

Proceedings of ISSI 2009, pages 165–172.
Edited by B. Larsen and J. Leta.

system based on concept profiles consisting of weighted Medical Subject Headings (MeSH) terms (Srinivasan, 2004) and the role of MeSH in ranking indirect connections in LBD was discussed by Swanson et al. in 2006 (Swanson etc., 2006). Bibliometric methods which were used for the semantic mapping of MEDLINE database in order to find unnoticed relations between articles of a discipline were analyzed by Spinak (Spinak etc., 1999).Seki and Mostafa proposed a formal model from information retrieval in order to discover implicit, hidden knowledge (Seki & Mostafa, 2007). Cory showed that Swanson's ideas are valid outside the biomedical field. He found an indirect link between a 20th century poet and an ancient philosopher in a humanities bibliographic database (Cory, 1997). An interesting application is proposed by Swanson, Smalheiser and Bookstein, namely fight against bioterrorism (Swanson etc., 2001).

In addition, literature-related discovery (LRD), including literature-based discovery (LBD) and literature-assisted discovery (LAD), was discussed by Kostoff (Kostoff etc., 2008a, 2008b), and a special issue presented the LRD methodology and voluminous discovery result from five problem areas: four medical (treatments for Parkinson's Disease (Kostoff etc., 2008c), Multiple Sclerosis (Kostoff etc., 2008d), Raynaud's Phenomenon (Kostoff etc., 2008e), and Cataracts (Kostoff ,2008f)) and one non-medical (Water Purification (Kostoff etc., 2008g)).

2.2 Related Research in China

The study on literature-based knowledge discovery in China was about ten years later than abroad. Generally, only two aspects of literature-based knowledge discovery were discussed in our country: introductions to literature-based knowledge discovery methods and simulating process of Swanson's knowledge discovery. As for the first aspect, methodological enlightenment and significance of Don R.Swanson's achievements in information science was discussed by Ma Ming and Wu Yishan (Ma, M. & Wu, YS., 2003); Enlightenment of development of traditional Chinese Medical Science from Dr. Swanson's literature-based knowledge discovery was concerned by Xu Jianyang et al.(Xu, JY., etc. 2005); An Xinying et al.(An, XY., & Leng, FH., 2006) compared and analysed four methods, word based lexical statistics method, phrase based lexical statistics method, concept based method and concept based lexical method of it; Zhang Yunqiu et al. (Zhang, YQ., & Leng, FH., 2008) introduced key techniques of literature-based knowledge discovery to readers in China. Considering the second aspect, Hao Liyun and Guo Qiyu (Hao, LY., & Guo, QY., 2007a, 2007b) carried out open knowledge discovery and practiced subject analysis method. Besides medical area, Arrowsmith, the tool of literature-based knowledge discovery, was used in aerospace area in China (Cao, ZJ., & Leng, FH., 2008).

3. Materials

The major database in our approach is Chinese Social Sciences Citation Index (CSSCI), and data of it were used as the sample in our research in order to demonstrate our method. The CSSCI project was undertaken in order to build up a scientific and fair research and evaluation system for Chinese social sciences (Su, XN. etc. 2001) and then Chinese Social Sciences Research Evaluation Centre (CSSREC) was established in 2000 in Nanjing University for the further development of CSSCI project.

Each article in CSSCI is associated with a set of keywords that describe the content of the associated discipline. Keywords are from an uncontrolled vocabulary and thesaurus used for indexing articles and, in particularly, for CSSCI database. With 65529 keywords, the sample of our research came from 528 journals and 124 second level disciplines in 2007 CSSCI database. The entire database searching in this paper was carried out in December 2008.

4. Methods

4.1 Model of LMKD

In Fig.1 we can see the model of LMKD. From inputting a set of keywords of serious articles of different disciplines, using a keywords extraction process, we build the keyword array of each discipline, cf. step 1. Each keyword array represents the knowledge of a special discipline. We represent this knowledge in the knowledge base in a formal form as a set of keywords, association rules between these keywords and additional background knowledge about the keywords. The association rules (based on keywords co-occurrence and described in section "keywords extraction") are used to denote the known relations between the keywords, cf. step 2. The discovery algorithm proposes new relations between the keywords based on the association rules (known relations) and disciplines of each keyword, cf. step 3. Therefore, the relation of each discipline is created, cf. step 4. Now we describe in more detail the basic components of this LMKD model.

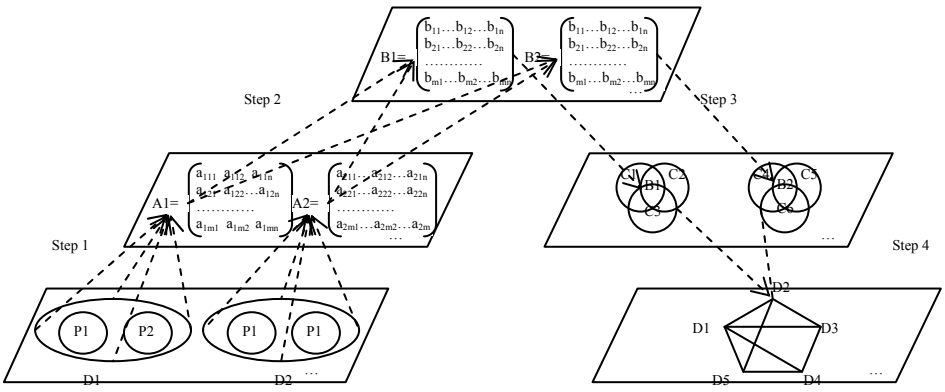

Fig 1. Model of Literature-based Multidiscipline Knowledge Discovery

4.2 Keywords Extraction

Creating keyword arrays for each discipline (Step 1)

In our discovery model, we add keywords from documents of each discipline to create a keyword array of every discipline, e.g. the keywords of two papers called "P1" and "P2" belonging to a discipline "D1" were extracted and a three dimensional keyword array called "A1" was created. In a considerable number of cases, one concept can have more than one keyword, but we considered these keywords as different concepts in our article for the following study on an academic vision of each discipline. Rows of A1 indicate different papers of a discipline, columns of A1 indicate different keywords of each paper, and three digits on the subscript of each element of A1 indicate discipline, paper and keyword from left to right. For example, a321 is the first keyword from the second paper of the third discipline. Then, keyword arrays "A2", "A3"… were created.

$$A1 = \begin{pmatrix} a_{111} & \cdots & a_{11n} \\ \vdots & \ddots & \vdots \\ a_{1m1} & \cdots & a_{1mn} \end{pmatrix} \quad A2 = \begin{pmatrix} a_{211} & \cdots & a_{21n} \\ \vdots & \ddots & \vdots \\ a_{2m1} & \cdots & a_{2mn} \end{pmatrix}$$

Creating relation array for each keyword (Step 2)
If the element "aij" of A1 is the element "amn" of A2, A1 and A2 are connected by "aij" and "amn".

$$A1: a_{ij} = A2: a_{mn} \rightarrow A1, A2 \text{ (connected)}$$

If the element "aij" of A1, the element "amn" of A2 and the element "axy" of A3 are the same keyword, this keyword is called "B1"; relation array B1 is created presenting the relation of A1, A2 and A3. Rows of B1 present different disciplines connected by B1 and columns of B1 present co-occurrence keywords of B1 in each discipline.

4.3 Discovery Process

Broadening academic vision (Step 3)
For a given starting keyword B (e.g. B1), first we find all the related keywords C (e.g. C1) for one discipline. In fact, elements of each row of relation array B are the terms of C, which is one dimensional keyword array. Obviously, terms of C are the academic vision of the discipline. Thus, we find all academic vision array C1, C2…Cn related to B1.
Next, we check if there are other same terms in C1 and C2 except for B1. If terms of C1 and C2 appear together, we found an existing academic vision between C1 and C2 (e.g. E1 in Fig 2.). If terms of C1 and C2 do not appear together, we discovered a potentially new academic vision between C1 and C2. The remaining keywords of C1 and C2 are now ranked and displayed to the user for evaluation and further investigation. It should be stressed that, in general, it is possible to have more than one intermediate keyword B1 on the path from C1 to C2.

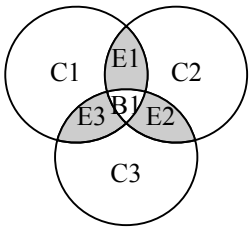

Fig 2. Academic Vision Connections among Disciplines

Finding relation among disciplines (Step 4)
Direct and indirect connections between every two disciplines have been found, the relations of those disciplines will be clearly visualized by visualization tools (e.g. Pajek, Ucinet).

4.4 Operating Procedures of LMKD

The method of LMKD must be easy to understand for researchers. So the operating procedures of it have been summarized as follows.

Table 1. Operating Procedures of Literature-based Multidiscipline Knowledge Discovery

No.	Operating Procedures
1	Let D (e.g. D1,D2) be given starting disciplines of interest
2	Find all keyword arrays A (e.g. A1,A2) such that there is an association rule D → A
3	Create all relation arrays B (e.g. B1,B2) such that there is an association rule A → B
4	Find all academic vision array C (e.g. C1,C2) such that there is an association rule B → C
5	Rank and display C to researchers
6	Find all relation between every two D such that there is an association rule C → D
7	Visualizing relation network of D to researchers

5. Results

Although LMKD method can be used for knowledge discovery support in general, we think it is especially useful for finding new academic visions and relations among disciplines. This is a consequence of the integration of background knowledge and is a unique feature not present in other literature-based discovery methods. There are several new possible application scenarios of our method.

In one scenario, we might start with a keyword for which two research regions are known, but the academic vision of two research regions are different. Through this intermediate keyword (e.g. harmonious society) we can try to find the difference between two research regions. The intermediate keyword should reveal something about the mechanisms of the influence of one discipline on the other discipline.

In a second possible application, two disciplines are known to be related as a result of association or linkage study, but the nature of their relationship is not known. Here we concentrate on the intermediate keywords, which should give us an idea about the relationship.

In the third scenario, LMKD can be used in a more widely research area than other literature-based discovery methods.

6. An Example

We analyzed the full 2007 CSSCI database as of the end of 2008. LMKD's operating procedure as follows.

- Procedure 1: 124 second level disciplines in CSSCI database were the given starting disciplines.
- Procedure 2: Find 124 keyword arrays for 124 disciplines. For example, keyword array of information science as follow.

$$\text{Information Science} = \begin{pmatrix} \text{information science} & \text{content analysis} & \text{word frequency analysis} \\ \text{knowledge network} & \text{citation network} & \text{knowledge map} \\ \cdots & \cdots & \cdots \\ \text{information retrieval} & \text{basic principle} & \text{knowledge management} \end{pmatrix}$$

- Procedure 3: According to 124 keyword arrays, 65529 relation arrays were created.

- Procedure 4: Based on relation arrays, academic vision was broadened. Supposing "harmonious society" is B1, 27 disciplines will be connected by it, e.g. philosophy, sociology, ethnics, religions, communication and law.
- Procedure 5: Academic vision of each discipline was ranked and displayed to researchers. For example, academic visions of sociology are "scientific outlook on development", "social security", "social services" and "social structure", but the academic visions of philosophy are "scientific outlook on development", "socialism", "values" and "fairness and justice". The difference between two disciplines is clearly listed.
- Procedure 6: When we discussed about "harmonious society", we could find that "scientific outlook on development" was the same vision of those two disciplines too. "Scientific outlook on development" connected other disciplines except the disciplines that "harmonious society" connected. So a complex relation network was established by the potential relationships among disciplines.
- Procedure 7: Relation network visualization.

7. Conclusion and Future Work

We present a new method of knowledge discovery: literature-based multidiscipline knowledge discovery (LMKD), which is quite different from the well known literature-based knowledge discovery method given by Don R. Swanson. It is a new application of bibliometrics, and can be used as a research idea generator to broadening academic vision. Because of the creating of academic vision arrays, it is easy to find the similarities and differences of academic vision among serious disciplines. Of course, the visualization of relation network is very helpful to create a relationship in researchers' mind.

We would like to highlight in this section some of the terminology problems we faced during the development of LMKD. Keyword field is the primary source of keywords in our LMKD approach and we were forced to detect and extract keywords from keyword fields. However, keywords from keyword field are not standardized, e.g. "WWW", "Web" and "Internet" are the same concept, but this concept was described in many different keywords given in different articles by different authors. So this problem should be resolved in future.

Acknowledgment

The authors would like to acknowledge the support from National Natural Science Foundation of China (70673071/G0309) and constructive comments from the reviewers.

References

An, XY., & Leng, FH.(2006). Study on Disjoint Literature-based Discovery. *Journal of the China Society for Scientific and Technical Information,* 25 (2), 87-93. (in Chinese)

Cao, ZJ., & Leng FH.(2008). Study on Literature-based Discovery in Aerospace. *Information Studies: Theories and Application,* 31 (4), 569-572. (in Chinese)

Cory, K.A. (1997). Discovering hidden analogies in an online humanities database. *Computers and the Humanities,* 31, 1-12.

Gordon, M.D., & Lindsay, R.K. (1996). Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. *Journal of the American Society for Information Science and Technology,* 47, 116-128.

Kostoff, R.N. (2008a). Literature-related discovery (LRD): Introduction and background. *Technological Forecasting and Social Change*, 75(2), 165-185.

Kostoff, R.N., Briggs M.B & Solka J.L. etc. (2008b). Literature-related discovery (LRD): Methodology. *Technological Forecasting and Social Change*, 75(2), 226-238.

Kostoff, R.N., & Briggs M.B. (2008c). Literature-related discovery (LRD): Potential treatments for Parkinson's Disease. *Technological Forecasting and Social Change*, 75(2), 226-238.

Kostoff, R.N., Briggs M.B. & Lyons T.J. (2008d). Literature-related discovery (LRD): Potential treatments for Multiple Sclerosise. *Technological Forecasting and Social Change,* 75(2), 239-255.

Kostoff, R.N., Block J.A., & Stump J.A. et al. (2008e). Literature-related discovery (LRD): Potential treatments for Raynaud's phenomenon. *Technological Forecasting and Social Change,* 75(2), 203-214.

Kostoff, R.N. (2008f). Literature-related discovery (LRD): Potential treatments for Cataracts. *Technological Forecasting and Social Change,* 75(2), 215-225.

Kostoff, R.N., Solka J.L., Rushenberg R.L. et al. (2008g). Literature-related discovery (LRD): Water purification. *Technological Forecasting and Social Change,* 75(2), 256-275.

Hao, LY., & Guo, QY.(2007a). Practice of Non-Interactive Literature-Based Knowledge Discovery through Subject Analysis (Ⅰ): Simulating Swanson's knowledge discovery process. *Journal of the China Society for Scientific andTechnical Information,*26 (5), 741-747. (in Chinese)

Hao, LY., & Guo QY.(2007b). Practice of non-interactive literature-based knowledge discovery through subject analysis (Ⅱ): Mining new knowledge in literatures on type 2 diabetes. *Journal of the China Society for Scientific andTechnical Information,*26 (6), 845-850. (in Chinese)

Ma, M., & Wu YS.(2003). Methodological Enlightenment and Significance of Don R.Swanson s Achievements in Information Science. *Journal of the China Society for Scientific andTechnical Information,*22 (3), 117-134. (in Chinese)

Weeber, M., Klein, H., De Jong-Van Den Berg, et al.(2001). Using concepts in literature-based discovery: simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries, J. Am. Soc. *Journal of the American Society for Information Science and Technology,52*(7), 548-557.

Weeber, M., Vos, R., Klein, H., De Jong-Van Den Berg, et al. (2003).Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide, *Journal of the American Medical Informatics Association,* 10 (3), 252-259.

Seki, K., Mostafa, J. (2007). Literature-based discovery by an enhanced information retrieval model, *Discovery Science, Lecture Notes in Artificial Intelligence.* 4755, 185-196.

Smalheiser, N.R., & Swanson, D.R. (1998). Using ARROWSMITH: A computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine,* 57, 149-153.

Spinak, E., Launy, S., & Grillo, B. (1999). Detection of unknown public knowledge through the semantic mapping by disciplines in large databases. *Proceedings of the 7th Conference of the international society for scientometrics and informetrics,* Colima (Mexico), 457-467.

Srinivasan, P. (2004). Text mining: generating hypotheses from MEDLINE, *Journal of the American Society for Information Science and Technology,* 55(5), 396-413.

Su, XN., Han, XM. & Han, XN. (2001). Developing the Chinese Social Science Citation index. *Online Information Review*, 25(6), p. 365-369.

Swanson, D.R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine,* 30, 7-18.

Swanson, D.R. (1987). Two medical literatures that are logically but not bibliographically connected. *Journal of the American Society for Information Science and Technology,* 38, 228-233.

Swanson, D.R. (1988). Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine,* 31, 526-557.

Swanson, D.R. (1989a). Verification of Results That Logically Related Noninteractive Literatures Are Potential Sources of New Knowledge. *Journal of the American Society for Information Science and Technology,* 40(3), 152-160.

Swanson, D.R. (1989b). A second example of mutually isolated medical literatures related by implicit, unnoticed connections. *Journal of the American Society for Information Science,* 40, 432-435.

Swanson, D.R., Smalheiser, N.R., and Bookstein A. (2001). Information discovery from complementary literatures: categorizing viruses as potential weapons. *Journal of the American Society for Information Science and Technology,* 52(10), 797-812.

Swanson, D.R., Smalheiser, N.R., & Torvik, V.I. (2006). Ranking indirect connections in literature-based discovery: The role of medical subject headings. *Journal of the American Society for Information Science and Technology,* 57(11), 1427-1439.

Xu JY, Ma, M., Wang, MK., et al.(2005). Enlightenment of development of traditional Chinese medical science from Dr. Swanson's non-interactive literature-based knowledge discovery. *Worl Science and Technology: Modernization of Traditional Chinese Medical and Materia Medica,* 7 (2), 48-54. (in Chinese)

Zhang, YQ., & Leng, FH.(2008). A Study on Key Techniques for Disjoint Literature-based Discovery. *Journal of the China Society for Scientific and Technical Information,* 27 (3), 521-527. (in Chinese)