Literature-based discovery: a classical approach of information science

Lin Wang[1], *Wei* Lin[2,*], and *Miaomiao* Zhao[3]

[1]China Academy of Science and Education Evaluation, Hangzhou Dianzi University, Hangzhou, China
[2]Tongji Medical School, Huazhong University of Science and Technology, China
[3]Independent Researcher, Tianjin, China

Abstract. The implicit links between documents are often of great significance to scientific discovery. This paper briefly reviews the background, basic ideas, and the related research tool Arrowsmith system of the literature-based discovery method founded by Professor Don. R. Swanson of the University of Chicago. It represents a classical research approach in information science. It is believed that this method has opened up a broader research field for information retrieval and knowledge discovery, provided new research ideas for medical informatics, and injected new vitality into information science.

Keywords: Literature-based discovery, Knowledge discovery, Arrowsmith, Swanson, D.R.

1 Introduction

Literature-based discovery provides an innovative research method for information professionals in the era of exponential growth of scientific information. It provides a new method and tool to for information retrieval and analysis. It not only directly facilitates the discovery of new knowledge, but also highlights the intermediary role of information professionals in scientific research. It represents a new horizon of library and information science.

The founder of literature-based discovery, information scientist Don.R.Swanson(1924-2012) developed the Arrowsmith system. The system provides new technical methods for effectively extracting useful knowledge from massive amount of information. It is a knowledge discovery platform that can solve the problem of knowledge fragmentation caused by discipline division and information explosion The Arrowsmith system has a great ability to explore the implicit relationship between knowledge, which has the possibility of being been approved by empirical studies in medical field. Generally speaking, the Arrowsmith system has a good prospect for knowledge discovery in medicine field.

* Corresponding author: linwei0602@qq.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

2 The review of literature-based discovery

With the rapid development of modern science and technology, the S&T knowledge structure evolved into a crisscrossed, highly complex network. During this process, the branches of each discipline are becoming detailed, which makes the relevant literature seem unrelated due to barriers between disciplines. If the potential knowledge links from a large number of scientific literature are discovered, it will promote the development of science [1]. As for the field of biomedicine, it has generated a tremendous amount of research data. If researchers can reveal the potential associations among the literature that are not directly connected then put forward and verify valuable scientific hypotheses through experiments, it will be great progress in the biomedical field.

In the 1980s, Professor Don R. Swanson of Chicago University proposed the literature-based discovery method. The basic idea of Swanson's literature-based discovery method is to establish the association between bibliographically unconnected literature sets A and C through the intermediary word set B. It is possible to dig out the invisible relationship between literature sets A and C that has not been publicly known by studying the correlation between AB and BC. Therefore, the implicit logical relationships among the research results in the scientific publications become visible. Since then, many scholars such as Smalheiser, Gordon, Lindsay, Wrerr, Padmini Srinivasan, Weeber, and Sebastian have improved the literature-based discovery method and promoted the popularization of this method.

literature-based discovery methods are classified into four types [2]: word-based frequency statistics, phrase-based frequency statistics, concept-based method, and concept-based frequency statistics. Among them, the phrase-based frequency statistical method is proposed by Gordon and Lindsay. It is an open knowledge discovery process that fully considers the semantic variants of root words and explores the relationship between concepts. The concept-based knowledge discovery method realizes the concept mapping between natural language and UMLS by employing the semantic types of UMLS to obtain the mutual relationship among co-current concepts and conduct knowledge mining. The concept-based word frequency statistical method uses the word frequency weight rather than simple word frequency. It takes concept and metadata as the basic research unit. Swanson's literature-based discovery method is the most representative of the word-based frequency statistical methods.

Generally speaking, the method of literature-based discovery has two steps: hypothesis formulation and hypothesis test [3]. The step of hypotheses formulation is also called the method of open, non-relevant literature knowledge discovery, which can be summarized as follows: C → B → A. That is to say, starting from the subject literature C, we can find literature set B that links literature set C and A together to establish the relationship between C and A. This kind of open literature-based discovery method is usually employed to find new treatments for diseases or new targets for drugs. It is necessary to set the threshold on the frequency and semantic type of the relevant term B and the target term A to screen the effective words and reduce workload. The step of hypothesis test refers to the closed non-relevant literature-based discovery. This method is usually used in hypothesis verification after researchers apply the open method to formulate a hypothesis. When the hypothesis that the implicit relationship between A and C exists is made, the closed knowledge discovery method can be applied to explore B from both ends of literature sets A and C. The researchers aim to get as many intermediary terms B related to both literatures sets A and C as possible by searching and processing literature to verify the hypothesis in detail. The more reasonable connection between A and C, the more valuable the hypothesis is. This process can be represented as A → B ← C [3].

3 The arrowsmith system

Swanson developed a human-computer interaction system using the MEDLINE database as the data source based on the literature-based discovery principle. He named it Arrowsmith. The Arrowsmith system has two versions (http://kiwi.uchicago.edu and new version http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html) . Compared with the previous version, the new version's interface is more friendly, and the functions are perfect: it breaks the limitation that the old Arrowsmith system can only be applied to study subject headings. The new version of the system can also be applied to retrieve and analyze both abstracts and subject headings. It integrates UMLS and realizes automatic sorting [4]. Due to the close connection with PubMed, the operation of the new Arrowsmith system becomes more accessible, faster, and flexible. The Arrowsmith system does not have its own database. It mainly makes it easier to find the invisible connection between the two documents sets by expanding the retrieval function of PubMed [5]. Arrowsmith relies on PubMed and other databases because all its data comes from the search results of MEDLINE and other databases. Therefore, the quality of PubMed search results will directly affect scientific discoveries of Arrowsmith. In essence, it is a knowledge discovery tool that expands PubMed retrieval functions to find implicit connections between the two sets of medical literature. If A and C are not relevant or only weakly related, searching "A and C" query in PubMed will result in little information about the relations between A and C. The Arrowsmith system expands the retrieval function of PubMed, so that the implicit association between A and C can also be retrieved and discovered [6]. Although the Arrowsmith system is valuable in knowledge discovery, verifying scientific hypotheses still depends on the lab experiments and research practice. The Arrowsmith system can not replace traditional empirical scientific research. However, it provides new clues for scientific research and helps researchers formulate reasonable hypotheses for innovation, which can greatly reduce the cost of medicine R&D [6,7].

Swanson has made several important discoveries by applying the literature-based discovery method and Arrowsmith system: the first in the medical field is that dietary fish oil has a certain effect on treating Raynaud's disease [8]. Studies show Raynaud's disease is a disorder of blood circulation. Most patients show typical symptoms of blood viscosity and increased platelet aggregation, and blood vessel constriction caused by blood and vascular disease. Meanwhile, it has been found that the active substance of dietary fish oil can reduce blood viscosity and platelet aggregation. Swanson used 34 research reports on Raynaud's disease caused by blood pathology as literature set A and 25 research reports on the amelioration of blood viscosity caused by dietary fish oil intake as literature set C. He then links the two sets of literature together through the list of Raynaud's disease subject headings set B. Swanson inferred from these results that dietary fish oil might be beneficial to treat Raynaud's disease. This scientific hypothesis was confirmed in clinical experiments two years later.

The second case is the study of the association between migraine headache and magnesium deficiency [9]. There are two main potential associations between magnesium and migraine headache: First, the low level of cerebral cortex function is one of the possible causes of migraine headache, and magnesium can effectively restrain the decrease of cerebral cortex function. Secondly, in clinical practice, the rate of magnesium deficiency is regarded as the diagnostic standard of epilepsy. The latter is related to migraine headaches. Prior to Swanson's study, there was no research exploring the possible association between magnesium and migraine headache. Swanson put forward the claim that dietary magnesium may alleviate migraines. After that, at least 12 clinical research reported that patients' migraine headaches were on the mend by dietary magnesium supplements. They generally showed a lack of magnesium in the whole body or part. Swanson's hypothesis about the

relationship between magnesium deficiency and migraine headache thus was supported. It should be mentioned that Swanson's these two knowledge claims (fish oil and Raynaud's disease, dietary magnesium and migraine headache) have also been verified in Chinese scientific literature [10,11].

Since then, Swanson and his colleagues conducted several other literature-based discovery studies by employing the Arrowsmith system [12,13,14,15,16]: magnesium deficiency and nervous system diseases, indomethacin and Alzheimer's disease, estrogen and Alzheimer's disease, free calcium phospholipase A2 and schizophrenia, and viruses that can potentially be used as biological weapons, atrial fibrillation and running. Other scholars also apply the literature-based discovery method to explore new medical ideas. Weeber et al. found new uses of the drug thalidomide for acute pancreatitis, myasthenia gravis, etc [17]. Srinivasan, Libbus and Sehgal showed that curcumin has a beneficial role in retinal diseases [18]. Due to Swanson's breakthrough contribution to methodology innovation of information science, the American Society for information science and Technology (ASIS&T) awarded him the Award of Merits in 2000.

4 The application of arrowsmith system in traditional Chinese medicine research

Since the literature-based discovery method and Arrowsmith system were introduced into China, many Chinese scholars have been interested in exploring the potential effect mechanisms of traditional Chinese medicine and repurposing it by applying the Arrowsmith system. For example, Li, Ge and Su followed the standard procedure of this method and applied the Arrowsmith in Chinese medicine. They found that some components of Chinese Angelica may be used to treat dysmenorrhea [19]. Wang and Xiao used the Arrowsmith to explore the detoxification mechanism in the combination of radix stephanine tetrandrae and aconiti lateralis preparata [20]. Chen, Zhang and Qin used the Arrowsmith system to discuss the correlation between cordyceps sinensis and the effect of Vitamin D . They proposed that these Chinese traditional medicines may generate the effect of treatment through Vitamin D [21].

Other scholars showed interest in the correlation between Chinese traditional medicine and depression. By employing Arrowsmith, Zhou. et al.investigated the possible mechanism of Radix Bupleuri in depression treatment. They argued that the effect of Radix Bupleuri might be related to grown factor, brain-derived neurotrophic factor [22]. Gao. et al.recently used Arrowsmith to study the mechanisms of Chinese medicine formulae in Song Dynasty- Xiaoyao Powder and its single herb in depression treatment [23]. As for intestine diseases, Tan et al. explored the latent mechanism of portulace oleracea.L in ulcerative colitis (UC) treatment. By analyzing the data retrieved from Arrowsmith, they found portulace oleracea L. may regulate NF-KB, TNF α , AKT, IL-6, thus having a curative effect on UC[24]. Tian et al. reported the modern science mechanism of fagopyrum cumosum and its effective components in treating irritable bowel syndrome (IBS) by utilizing a literature-based discovery method. They may regulate toll-like receptors, TNF α , VEGF, and COX-2, the abnormal expression of which participate in IBS[25]. Based on the search results of Arrowsmith, Sun. et al. explored the curative mechanism of a Chinese patent medicine -- Tianfoshen Oral Liquid for colorectal cancer [26]. They argued that this Chinese patent medicine has the functions of down-regulating the expression of TLR4 and COX-2 in colorectal tissues, reducing abnormally high expression of VEGF, and facilitating the expression of apoptosis genes.

5 Conclusion

"Undiscovered public knowledge" [8] is a promising field that information scientists and professionals can show their unique strengths. As a pioneer, professor Don.R.Swanson of the University of Chicago developed an innovative knowledge discovery method called literature-based discovery, which has become a powerful research tool of information science for exploring the potential knowledge links between medical documents. His successful application of literature-based discovery and Arrowsmith system evidenced the indispensable and key role of information science in ordering and managing Popper's world 3, the objective knowledge world. Information professionals are not only the curator of objective knowledge resources, but also active knowledge creator and scientific discovery and innovation vanguard. This paper briefly reviews the background, basic ideas, and the related research tool Arrowsmith system of the literature-based discovery method. It also discussed the application of Arrowsmith system in Chinese traditional medicine field. Literature-based discovery represents a classical research approach in information science. It is believed that this method has opened up a broader research field for information retrieval and research, provided new research ideas for medical informatics, and injected new vitality into information science.

References

1. Ma M and Wu Y S 2003 Methodological enlightenment and significance of Don R.Swanson's achievements in information science. *Journal of the China Society for Scientific and Technical Information,* 22(3), pp259-266. (in Chinese)
2. An X Y and Leng F H 2006 Principles of non-interactive literature-based knowledge discovery. *Journal of the China Society for Scientific and Technical Information,* 25(1), pp. 87-93. (in Chinese)
3. Gao H J and Zhao K Y 2010 Knowledge Discovery Based on Non-interactive Literature Research Progress. *Journal of Liaoning University of Traditional Chinese Medicine* 12(6), pp. 133-135.(in Chinese)
4. Zhang Z J and Wang Y Y 2015 Kidney deficiency-phlegm and blood stasis-poison brewing-disease collaterals—understanding of early pathogenesis of senile dementia in traditional Chinese medicine. *Journal of Basic Chinese Medicine,* (3), pp.244-246. (in Chinese)
5. Ma M, Zheng C J and Xu J Y 2004 MEDLINE-based non-interative literature-based discovery. *Chinese Journal of Medical Library and Information Science,*13(5), pp.1-3 (in Chinese)
6. Dong F H and Lan X Y 2004 Document-based knowledge discovery tools—Arrowsmith. *Journal of Intelligence,* 23(5), pp.52-54. (in Chinese)
7. Guo W N and Si L 2013 Overview of research on non-interactive literature-based knowledge discovery in China. *Documentation, Information&Knowledge,* (06), pp. 97-105. (in Chinese)
8. Swanson D R 1986 Undiscovered public knowledge. *Library Quarterly* 56(2), pp. 103-118.
9. Swanson D R 1988 Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine,* 31(4), pp.526-557.

10. Huang S Q, Cheng C and Li Z Y 2008 Application of the open knowledge discovery method for non-interrelated literature in Chinese literature. *Information Studies:Theory&Application,* (02), pp. 246-250. (in Chinese)
11. Qian Q, Hong N, Li Y and An X Y 2012 Construction of the Chinese disjoint literature-based knowledge discovery system CmedLBKD. Information Studies:Theory&Application, 35(04), pp.109-113. (in Chinese)
12. Swanson D R 1990 Corroboration of migraine-magnesium connection. *Journal of the American Society for Information Science*, 41(4), pp. 310.
13. Smalheiser N R and Swanson D R 1996 Indomethacin and alzheimer's disease. *Neurology,* 46(2), pp. 583-583.
14. Smalheiser N R and Swanson D R 1996 Linking estrogen to alzheimer's disease: An informatics approach. *Neurology,* 47(3), pp.809-810.
15. Smalheiser N R and Swanson D R 1998 Calcium-independent phospholipase A2 and schizophrenia. *Archives of General Psychiatry,* 55(8), pp. 752.
16. 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*, 51(14), pp. 797-812.
17. Weeber M, Vos R, Klein H, Aronson A R and Molema G 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), pp.252–259.
18. Srinivasan P, Libbus B and Sehgal A K 2004 Mining medline: postulating a beneficial role for curcumin longa in retinal diseases. In Workshop BioLINK, Linking Biological Literature, Ontologies and Databases at HLT NAACL, pp.33–40.
19. Li W L, Ge Y L and Su S L. et al. 2008 Correlation between Angelica Sinensis and dysmenorrheal studied using knowledge-discovery tool Arrowsmith. *Chinese Journal of Medical Library and Information Science*,17(4), pp. 7-11. (in Chinese)
20. Wang R J and Xiao W 2016 Latent mechanism of detoxification in combination of radix stephaniae tetrandrae and radix aconiti lateralis preparata using knowledge-discovery tool Arrowsmith. *World Science and Technology: Modernization of Traditional Chinese Medicine*, 18(3), pp.527-531. (in Chinese)
21. Chen Y Z, Zhang J P and Qin Z et al. 2010 Potential correlation between the efficiency of cordyceos sinensis and vitamin D: Arrowsmith-based study. *Chinese Journal of Medical Library and Information Science*,19(12), pp.48-50. (in Chinese)
22. Zhou Y, Chen W and Chen Y Z et al. 2017 Discussion on radix bupieuri in the treatment of depression based on Arrowsmith. *Lishizhen Medicine and Materia Medica Research*, 28(05), pp.1250-1252. (in Chinese)

23. Gao Y, Xu T and Zhou Y Z et al. 2019 Study on mechanism of Xiaoyao powder anti-depression based on Arrowsmith tool. *Chinese Traditional and Herbal Drugs,* 50(14), pp.3484-3492. (in Chinese)
24. Tan Y Y, Zhan S M, Ding K and Ding Y J et al. 2017 Correlation between portulaca oleracea l. and ulcerative colitis by using knowledge-discovery tool Arrowsmith. *Journal of Liaoning University of Traditional Chinese Medicine*, 19(03), pp.118-121. (in Chinese)
25. Tian C, Yan J, Fan F T, Sun Z G, and Lu Y 2014 Correlation between fagopyrum cumosum and its effective components and irritable bowel syndrome studied using knowledge-discovery tool Arrowsmith. *Pharmacology and Clinics of Chinese Materia Medica,* 30(06), pp.190-193. (in Chinese)
26. Sun L H, Wang S L and Cao Y Z 2016 Investigate Tianfoshen oral liquid and correlation between Its effective components and colorectal cancer using knowledge-discovery tool Arrowsmith. *Chinese Journal of Experimental Traditional Medical Formulae,* 22(05), pp.215-220. (in Chinese)