Study on theoretical principles of noninteractive literature-based discovery

ZHANG Yunqiu[1] & LENG Fuhai[2*]

[1]School of Public Health, Jilin University, Changchun 130021, China

[2]Department of Information Research, National Science Library, Chinese Academy of Sciences, Beijing 100190, China

Abstract

Some studies find that noninteractive literature-based discovery has three theoretical principles, including retrieval theory, bibliometrics and logics. Thereinto, targeted retrieval strategy, co-occurrence theory, and syllogism are practical theories. The general principle of targeted retrieval strategy is narrowing range and improving relevancy and based on incomplete formalization of syllogistm, noninteractive literature-based discovery should change its research direction with a target of full automation and achieve optimization of filtration and sorting method in high-level concurrence framework, thus, form a knowledge discovery system with application value under human intervention.

Keywords

Noninteractive Literature-based discovery, Retrieval strategy, Co-occurrence, Syllogism

1 Introduction

Professor Swanson at the University of Chicago, United States, put forward noninteractive literature-based knowledge discovery in 1986. It was pointed out by Swanson that if independently created fragments of knowledge were logically related, brought together and interpreted, they could reveal useful information, which could be called undiscovered public knowledge (Swanson, 1986b). Those literatures containing undiscovered public knowledge were logically noninteractive or complementary but disjointed. The follow-up domestic articles named those as noninteractive literature (Ma & Wu, 2003). The intelligence research method from noninteractive literatures to identify effective, novel, useful and understandable knowledge was noninteractive literature-based knowledge discovery, which could help researchers find potential connection and boost generation of new knowledge. At present, though there has made some progress in finding and identifying automation approach, the argument on the fundamental theory can only be found in the article by Swanson in 1986, which explained the philosophical basis of noninteractive literature-based knowledge discovery by using Popper' world 3 (Swanson, 1986a). The article made some preliminary analysis on the theory basis on studying process and characteristics of noninteractive literature-based knowledge discovery.

* Correspondence should be addressed to Leng Fuhai (E-mail: lengfh@mail.las.ac.cn)

2 Retrieval theory

Literature retrieval is a process to gain the queried articles from vast amounts of document without major omissions in a quick and precise way according to the specific needs of the project by using scientific methods and applying specific tools (Lai et al., 2006). The result is to gain the related documents. The connection can be understood from different angles. The connection in the article is referred as subjective, which means the subject or core content of the document cluster from retrieval system can be matched with user's information queries. As to bibliographic database, it is the process to extract the feature item of document content by the similarity comparison of article authors, title and descriptors, etc. According to such criterion, the retrieved documents generally have words or phrases marking the same subject, or same authors. During the process of noninteractive literature-based knowledge discovery, it is the most basic step, in fact, to construct the initial document cluster on certain subject by analyzing its internal and external features. In addition, it's necessary to conduct cross retrieval test through two subjects in the end of open noninteractive literature-based knowledge discovery and the beginning of enclosed noninteractive literature-based knowledge discovery, which is to identify the relatedness of the two subjects. Only when the two documents clusters have no cross references, there has the possible meaning to detect noninteractive literature-based knowledge discovery. From the above analysis, noninteractive literature-based knowledge discovery can be regarded as expanded and complementary related document retrieval. Relatedness retrieval is the premise and the retrieval theory is one of its fundamental bases.

2.1 Retrieval strategy

Retrieval strategy is a thorough plan and program made for retrieval goal, which can design and guide the whole retrieval process. The conventional strategies are raised by Charles Bourne, such as most specific facet first, lowest postings facet first, build-block, citation pearl-growing and successive fractions, etc. As for the current search system, retrieval strategy, in its narrow sense, is the construction and arrangement of retrieval expression, whose essence is to analyze the features of database record structure and retrieval subjects.

2.2 Retrieval features

As for the ABC model of noninteractive literature-based knowledge discovery, it is directional no matter whether the discovery process is open or enclosed. So the authors think that it is a targeted retrieval strategy in the noninteractive literature-based knowledge discovery. The so-called target retrieval strategy, in the noninteractive literature-based knowledge discovery, is to make a whole retrieval plan and program according to the expected discovery type. As shown in figure 1, in the

open ABC model, the range is increasing gradually from initial cluster A to intermediate cluster B, then to target cluster C. The subset of A, B and C, A', B' and C' on the dotted figure line are discovery directing the target concept. Therefore, the general principle of target retrieval strategy is to narrow the range, raise relatedness with the subject and then increase retrieval accuracy. Target retrieval strategy defines the improvement of filtering and sorting method in the process of noninteractive literature-based knowledge discovery.

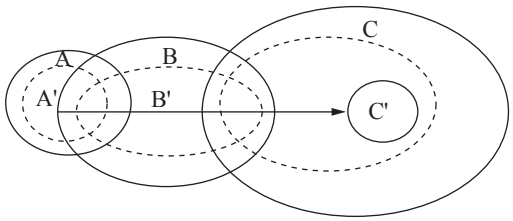


FIG. 1. Targeted retrieval strategy

3 Bibliometrics theory

Bibliometrics is a branch subject of library and information science (Egghe & Rousseau, 2002). It utilizes quantitative analysis and statistics to describe, assess and predict the present conditions and development tendency of science and technology by a variety of literature characteristics. Bibliometrics used in noninteractive literature-based knowledge discovery mainly focuses on the co-occurrence. Firstly, the hypothesis as noninteractive literature is no citation, cited and co-citation; secondly, term co-occurrences (high-order word co-occurrence) make it feasible in the discovery process. Thus co-occurrence is one of the theory bases on noninteractive literature-based knowledge discovery.

3.1 Co-occurrence theory

Co-occurrence in the scientific papers refers to same or different property of two things happening at the same time, which ranges from co-occurred subjects (title words, key words, subject words, etc.) happening among pieces of articles, co-authors, co-institution, paper and key words to authors-institution co-occurrence, etc. Co-occurrence can be listed as document coupling, co-citation, co-paper, term co-occurrence, author co-citation, and journal co-citation (Yang, 2006). Co-citation and term co-occurrence are mainly applied in noninteractive literature-based knowledge discovery. Co-citation has the features that two documents are cited by others at the same time. Generally, co-cited documents have more or less subject similarity. Co-citation numbers, that is co-citation strength, can be used to measure relatedness between two documents. If two or more documents are co-cited, it's possible to find them logically complementary. If

consider them together, one hypothesis may be raised, or find one solution for a problem. So at the initial stage, it is most important to remove co-citation in noninteractive literature-based knowledge discovery. Term co-occurrence are words representing document—that is, key words, title words or subject words, and so on, appear in the same document, through which different document can be linked to each other. It is the most basic thought in noninteractive literature-based knowledge discovery.

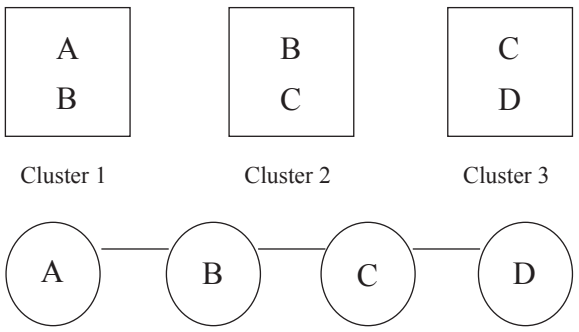
3.2 Co-occurrence features

3.2.1 Term co-occurrence

At present, noninteractive literature-based knowledge discovery mainly centers on same kind of characteristic items representing document subject, that is, term co-occurrence. In the bibliography database, the co-occurred words are derived from title, abstract and subject terms. The term co-occurrence that indicates here differs greatly with co-word analysis, which is employed, by measuring development track, characteristics, and the connection strengths among fields or disciplines through the counted words association, usually key words, to reveal R&D level, trends and general micro knowledge structure in a specific field, to analyze field disciplines and static structure in the lateral and longitudinal way and to expand IR fields so as to help users retrieve information. In contrast to noninteractive literature-based knowledge discovery, a set of co-occurring terms is extracted under the certain conditions that the related documents is clustered around one subject and the initial concept is clear. The clustered co-occurring terms set describes the word context, analyzes the relatedness strength between the co-occurring word and the initial word, and then finds out possible potential connection. From above mentioned content, we can see that there exist different focus, purpose and the applied analysis technology between the co-occurring word in noninteractive literature-based knowledge discovery and co-word in bibliographics.

3.2.2 Second-order concurrence

Co-occurrence has different orders. If it is like figure 2, it is higher-order concurrence. As shown in figure 2, Swanson based noninteractive literature-based knowledge discovery on second concurrence. Second concurrence can be described as: concept A (initial concept) and B (intermediate concept) co-occurs in one or more documents (cluster A), concept B and concept C co-occurs in document cluster B. However, as for a new discovery, concept A and C must not co-occur in the whole documents because the term representing the subject in the paper is the symbol of scientific research content. So noninteractive literature-based knowledge discovery is currently based on second concurrence, which can also be expanded to other representing items, such as author and institution.



First order concurrence: AB, BC, CD
Second order concurrence: AC, BD
Third order concurrence: AD

FIG.2. High-order concurrence.

Second-order concurrence model in the noninteractive literature-based knowledge discovery shows the following characteristics: ①Directional. In the framework of second concurrence, it is directional in the process of noninteractive literature-based knowledge discovery. The key is an attempt to find out the meaningful connection. Among different types of initial concept, the target concept is totally different. So the semantic relatedness of concept is vitally important. ②Strength of association. In the noninteractive literature-based knowledge discovery, different potential association can be found according to the different strength of association, which is also the key factor to detect the final result. The term co-occurrence is feasible in the discovery because the final potential connection is judged in fact by argument in the noninteractive literature-based knowledge discovery. Current method is to analyze the title, abstract and mesh items, sometimes only mesh items. However, the mesh only describes the subject concept of the whole document in general without deepening into semantic sentence, which is controversial, therefore the method by mesh only has shortcomings in the noninteractive literature-based knowledge discovery. While turning to another aspect of mesh, you may find that mesh is a kind of normative words revealing the document concept, which can play a connection role in analyzing co-occurrence word representing other subject items. It is one main factor to measure strength of association.③Connection transitivity. As shown in the figure 2, the direct connection of higher-order concurrence can be transitive. April Kontostathis and William M. Pottenger (2006) tried to reveal latent semantic connection of high-order concurrence including second and third concurrence by using latent semantic indexing (LSI). The results indicated that the latent semantic connection of second concurrence is stronger than third concurrence; meanwhile, the semantic relationship has transitivity. From the above analysis we can see second concurrence model has its theory basis applied in noninteractive literature-based knowledge discovery.

4 Logic theory

Logic is a study of reasoning for which is true and is valid from the form or structure. Reasoning is a thought process from the known knowledge as premise to deduct new knowledge as conclusion. Noninteractive literature-based knowledge discovery is in fact a reasoning process of deducting valid, novel, potentially useful and understandable knowledge from known knowledge as premise. Hence, logic may be looked as one of theory basis.

4.1 Syllogistic logic reasoning

Syllogistic logic is the deductive reasoning analysis of the judgments into propositions which is involving two judgments related by a common concept (Copi & Cohen, 2007). It can be seen that two judgments are related by means of "common concept" to constitute scientific reasoning form. Syllogistic logic reasoning is consisted of three arguments, of which two are the premises, one is conclusion. Aristotle's syllogistic logic centers on discussing the relatedness between the premise and conclusion in form, clarifies what kind of logic form between two premises and one conclusion have such relatedness. The lattice, syllogistic logic structure, is determined according to the intermediate argument, and then the modal, syllogistic logic reasoning forms, by the logic form of three arguments consisting of premises and conclusion. Thus with the determined form, structure and a "have-to-be" variety of different forms of syllogistic logic reasoning, there came into valid reasoning form and developed a system containing these reasoning forms (Copi & Cohen, 2007).

In the noninteractive literature-based knowledge discovery, the argument can be clarified through one model for logical related literature and logical connection. Suppose there are two documents, in the first document, under certain circumstances, A could lead to B. For example, medicine A can change hormone B level of blood. The cause-effect relationship can be labeled as "AB"; there are similar cause-effect relationship "BC". For example, hormone B can affect the pace of disease C. Then it can be inferred that anyone who knows the premise of AB and BC may realize A could affect C, labeled as "AC". This syllogistic structure may be informally viewed as the marking of more sophisticated model of scientific arguments. So the ABC discovery model of noninteractive literature exactly put syllogistic logic reasoning into use.

4.2 Syllogism features

In the noninteractive literature-based knowledge discovery, it makes good use of syllogistic logic reasoning forms and focuses on the logic related content analysis. The author here illustrates the logic basis of noninteractive literature-based knowledge discovery through case study. Table 1 shows the logic structure prototype of eleven related pairs in the discovery of migraine and magnesium, among which A means magnesium; C means migraine; B means one or more

intermediate pathophysiology connection; Bn is the n-connections; "→"is "possible to affect".

From the view of form, the simplest model, A→B, B→C, can deduce A→C. The second, sixth, eighth and eleventh connection pair in table1 belongs to this model. And then B→A and B→C may deduce A→C or C→A. The first belongs to this model. In addition, A→B, B =C, so A→C; or A =B, B→C, then A→C. Here "=" means "actions are equal" or "mechanism is equal", also can be"equivalent". The third, fourth, fifth and tenth belong to this model. The remaining is B→Bn, Bn→C, that is, B→Bn→C. Considering about A→B, then A→C. The seventh and ninth belong to this model.

From the above deducing process of noninteractive literature-based knowledge discovery, it can be seen that there is no limitation in syllogistic forms. Those influences can be formed through a variety of indirect connection, a rather than a only way. Although some related structure is obviously simple likely in accordance with logic form reasoning process, it seems not reliable to easily judge that two things have certain connection only from the form. No matter it is from syntax or from semantic structure, it can't offer direct cause-effect deduction chain. Each segments of deduction on certain relatedness shall rely more on the understandable background knowledge. In the noninteractive literature-based knowledge discovery, it is not necessary to assume that formal connection must be transited. The focus is the hint, that is, logic of suggestibility, whose aim is to stimulate forming new possible assumption. Namely, when there happens to be two arguments, A→B and B→C, there comes one guess that A may affect C right now. In general, AB or BC arguments are hard to form conclusion. Most of the connections are also impossible to represent the only reason. Since any AC deduction is usually established by assuming to a certain degree, any final conclusion must be formed by human intervention.

TABLE 1. Logical structure of potential relatedness in the literatures of "migraine and magnesium" (Swanson, 1988).

Relatedness Pairs	ABC Symbols Defined	a) Symbols Defined	b) Symbols Defined	Potential Relatedness
a) Stress and Type A behavior are associated with migraine 1000-seed weight. b) Stress and Type A behavior lead to body loss of magnesium.	A: magnesium B: Stress and Type A C: migraine	B→C	B→A	A→C
a) Excessive vascular tone and reactivity may increase susceptibility to migraine. b) Magnesium can reduce vascular tone and reactivity	A: magnesium B: vascular tone and reactivity C: migraine	B→C	A→B	A→C
a) Calcium channel blockers have been used successfully in preventing migraine attacks. b) Magnesium is a natural calcium channel blocker.	A: magnesium B: Calcium channel blockers C: migraine	B→C	A=B	A→C

(Continued)

(Continued)

Relatedness Pairs	ABC Symbols Defined	a) Symbols Defined	b) Symbols Defined	Potential Relatedness
a) "Spreading cortical depression" is thought to be implicated in the early phase of a migraine attack. b) High levels of magnesium in the extracellular cerebral fluid inhibit spreading cortical depression in laboratory animals.	A: magnesium B: Spreading cortical depression C: migraine	B=C	A→B	A→C
a) There is evidence for a connection between epilepsy and migraine. b) Magnesium deficiency may increase susceptibility to epilepsy.	A: magnesium B: epilepsy C: migraine	B=C	A→B	A→C
a) Migraine patients have abnormally high platelet aggregability. b) Magnesium can suppress platelet aggregation.	A: magnesium B: platelet aggregability/ aggregation C: migraine	B→C	A→B	A→C
a) Platelets of migraine patients are abnormally sensitive to serotonin release. b) Magnesium can inhibit serotoniu-induced contractions of vascular smooth muscle.	A: magnesium B: serotonin B2: vascular contractions C: migraine	B→B2→C	A→B2	A→C
a) Substance P may be a cause of head pain in migraine. b) Magnesium has an antagonistic effect on substance P activity.	A: magnesium B: Substance P C: migraine	B→C	A→B	A→C
a) Abormal prostaglaudin (PG) release can aggravate vasoactivity in migraine. b) Magnesium increases prostacyclin (PG12) formation.	A: magnesium B: prostaglandin B2: vasoactivity C: migraine	B→B2→C	A→B	A→C
a) Migraine may involve sterile inflammation of the cerebral blood vessels. b) Magnesium has anti-inflammatory properties.	A: magnesium B: inflammation C: migraine	B=C or B→C	A→B	A→C
a) Cerebral hypoxia may play a key role in migraine. b) Magnesium can protect against brain damage from hypoxia.	A: magnesium B: hypoxia C: migraine	B→C	A→B	A→C

From the above analysis, it can be concluded that it is not advisable and practical to take total formal reasoning. So it is worthy of discussing the completely automated knowledge discovery in the noninteractive literature. Additionally, it is necessary for noninteractive literature-based knowledge discovery to add certain artificial participation in syllogistic logic reasoning process.

As a conclusion, according to the key steps of noninteractive literature-based knowledge discovery, here raises retrieval theory, bibliometric theory and logic theory as its theory base. These three theories offer theoretical guidance to raise the discovery efficiency among noninteractive literatures and endeavored direction to expand application fields. Based on incompletely syllogistic logic reasoning formalization, we shall shift the research direction of completely automated

knowledge discovery among noninteractive literatures, focus on filtering and sorting method optimization in the framework of high-order co-occurrence and under the guidance of target retrieval strategy, and then establish more practical knowledge discovery system with the aid of artificial participation.

Acknowledgement

The authors wish to thank two sponsors. One is "The theories, methods and applications on noninteractive literature-based knowledge discovery (07JA870005)" sponsored by Social Science Research Fund Planning Project from Ministry of Education (MOE), and the other is "The theories and methods research based on evolution and evidence of S&T Innovation (70873123)" by the Nature Science Foundation of China. We are also grateful for helpful suggestions from a number of professors for their valuable comments on an earlier draft of this paper.

References

Copi, I.M., & Cohen. (2007). *Introduction to Logic* (逻辑学导论). J. J. Zhang, et al. (Trans). Beijing: China Renmin University Press, China.

Egghe, L., & Rousseau, R. (2002).Co-citation, bibliographic coupling and a characterization of lattice citation networks. *Scientometrics*, 55 (3), 349- 361.

Kontostathis, A., & Pottenger, W.M. (2006). A framework for understanding Latent Semantic Indexing (LSI) performance. *Information Processing & Management*, 42 (1), 56 - 73.

Lai, M.S., Zhao, D.Q., Han, S.L., & Wang, Y.F. (2006). *Information retrieval by computer* (计算机情报检索). Beijing: Peking University Press, China.

Ma, M., & Wu, Y.S. (2003). Methodological enlightenment and significance of don r.swanson's achievements in information science (Don R. Swans的情报学学术成就的方法论意义与启示). *Journal of the China Society for Scientific and Technical Information* (情报学报), 22 (3), 259 - 266.

Swanson, D.R. (1986a). Fish oil, reynardps syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30 (1), 7 - 18.

Swanson, D.R. (1986b).Undiscovered public knowledge. *Library Quarterly*, 56, 103 - 118.

Swanson, D.R. (1988). Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31 (4), 526- 557.

Yang, L.Y. (2006). *The theoretical and applied study of occurrence and co-occurrence phenomena* (科技论文共现理论研究与应用). Beijing: Graduate University of the Chinese Academy of Sciences.

——Translated from *Journal of Library Science in China* (中国图书馆学报), 2009, no.4