A discussion of the bi-directional ranking of occurrence-frequency based non-interactive literature method for knowledge discovery①

ZHANG Yunqiu* & GUO Kelei

School of Public Health, Jilin University, Changchun 130021, China

Received Dec. 6, 2009
Accepted Dec. 21, 2009
Translated with a permission from *Information Science* (in Chinese), 2009(8):1240

Abstract Based on the analysis of the existing ranking terminology or subject relevancy of documents methods through an intermediary collection as a catalyst (designated as Group B collection) for the purpose of of non-interactive literature-based discovery, this article proposes a bi-directional document occurrence frequency based ranking method according to the "concurrence theory" and the degree and extent of the subject relevancy. This method explores and further refines the ranking method that is based on the occurrence frequency of the usage of certain terminologies and documents and injects a new insightful perspective of the concurrence of appropriate terminologies/documents in the "low occurrence frequency component" of three non-interactive document collections. A preliminary experiment was conducted to analyze and to test the significance and viability of our newly designed operational method.

Keywords Non-interactive literature-based knowledge discovery, B collection, Frequency of terminology occurrence

1 Introduction

The idea of a non-interactive literature based knowledge discovery was first introduced by Don R. Swanson, professor of University of Chicago in 1986[1]. Non-interactive literature means that one particular sphere of knowledge has no common ground in their subject tenets with another knowledge sphere. They do not cite each other, nor have been co-cited together by other scholars. It implies that there is no logical or intellectual relationship between these two spheres of knowledge. However, they may have the potential in getting connected together conceptually someday by a scholar engaging in his or her own knowledge seeking or knowledge creation undertakings[2]. The heated academic discussions of our time about the non-interactive literature-based knowledge discovery method suggest that a new alternative information research method is possible. This new research method is

① This work is supported by Humanities and Social Science Foundation of Ministry of Education of China (Grant No. 07JA870005).

* Correspondence should be addressed to Zhang Yunqiu (E-mail: zhangyq@mail.las.ac.cn).

CJLIS
Vol. 2 No. 4, December 2009
pp 31–42
National Science Library, Chinese Academy of Sciences

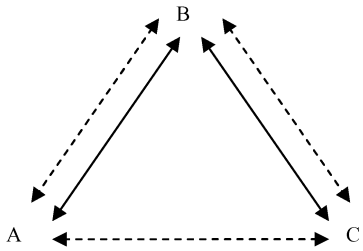

Fig. 1 Swanson's ABC discovery model.

deemed by many researchers to be more functionally effective, original, and full of potential in identifying a piece of new knowledge[3].

Most non-interactive literature-based knowledge discovery methods utilize Swanson's ABC discovery model (Fig. 1). This model has now been developed into either an open or closed model in its application on furthering knowledge discoveries.

An open model for knowledge discovery is derived from a hypothesis stemming from a knowledge discovery process, starting from a topic of scientific significance or from a research inquiry, as the starting point A (as shown in the above triangle-shaped graph, Fig. 1). Next, this conceptual point of inquiry is used as a search query to download or to retrieve desired information from either an online database and/or from all print documents available for this knowledge discovery task. These retrieved documents constitute the initial group of literature, which is essential for carrying on to the next step of research inquiry and is labeled as the literature of "situational grouping point A" (thereafter abbreviated as "Group A" or Group B" or "Group C"). Important terms or phrases are to be extracted and processed initially from this Group A. Each and every word or phrase resulting from a filtered list from Group A is termed as a set of intermediate research query instruments of "Group B."

As a selected group of research queries, these selected query terminologies represent comprehensively the most important and relevant subject queries embedded in "Group A" after consulting the same online database and the print collections again and by using the newly established queries extracted and to proceed to work on "Group B". The resulting set of documents retrieved from Group B is the second-tier intermediate group of literature, named as "Group B." This intermediate literature of "Group B" is then screened again in search for extracting a set of even more appropriate (subject relevant) terminologies and/or phrases necessary for moving the research project on hand forward. After a successful screening process, one particular group of query terms is finally selected as the most desirable instrument for the targeted research inquiry and possibly for the exploration of some new spheres of knowledge that may be visibly in the horizon. Finally, one can form a hypothesis that an appropriate number of research queries

distilled in "Group C" may show implicitly a logical or intellectual relationship with "Group A" via "Group B". Such a terminology/document screening and selection process is used to describe the "closed model" of knowledge discovery.

It starts with the assumption that there is an intellectual relationship of "Group A" and "Group C". The presumed connection between "Group A" and "Group C" may be just a newly discovered association or a previously assumed hypothesis. The new knowledge is gained by employing the "closed knowledge discovery model," which involves finding the appropriate queries in "Group B" as the possible intermediary catalyst capable of triggering the discovery or a validation of an assumed subject relationship. This knowledge discovery process is actually stemming from the instigation of two types of entities, namely, conceptual perceptions and published literature/documents.

For either the "open discovery model" or the "closed discovery model", constructing the Group B collection is the key step in the process of discovering a new knowledge by linking two pieces of unrelated knowledge spheres. The quality of Group B collection will affect the efficiency of knowledge discovery directly so much so many researchers are paying their undivided attention to this phenomenon. In this method of knowledge discovery, selecting relevant query terminologies/documents in the B Group is the most important task since a portion of the query terminologies/documents selected from Group B is to form the basis for searching and identifying the possible subject connection to those in "Group C."

As for how to select these research query terminologies/documents properly in Group B is a key issue to be dealt with. The most commonly used method is ranking the selected query terminologies in the Group B according to certain criteria. However, what kind of criteria that can reveal better the relevance of the subject matter is of the most importance. This article proposes a bi-directional occurrence-frequency based ranking method according to the "concurrence theory" and the degree and extent of subject relevancy. This method explores the ranking method based on the frequency of relevant terminology occurrence. More importantly, such a ranking method is to take into account of those terminologies of concurrence and in low frequency occurrence component into consideration. A preliminary experiment was conducted to test the significance and feasibility of this innovative method for the discovery of a new knowledge sphere.

2 The status quo of current research and problems

At present, the common ranking methods for relevant terminology are those occurrence-frequency based ones. They are basically consisted of two ranking methods, namely, 1) The simple frequency-based ranking method and 2) the averaging of frequency-based terminology concurrence ranking method. Simple frequency-based terminology ranking method extracts the terminologies or phrases

of Group B, which may also have such terminology concurrence in Group A or Group C. These methods rank the concurrent terminology that existed in any two or all those three groups in a descending order. Finally, the targeted terminologies from Group B are identified according to certain threshold value. The computing method for the threshold value is $n=(-1+\sqrt{1+8I_1})/2$. Averaging of frequency-based terminology concurrence ranking method refers to the occurrence frequency of a certain terminology in the document records of Group A or Groups C.

It is possible to use the "averaging of frequency based terminology concurrence ranking method" to eliminate the problem of those terminologies despite of their having a high frequency of occurrence yet not being able to ascertain the subject relevance of the documents retrieved. For instance, some terminologies do appear in high frequency in the documents covered for research. However, since they are of such a broad concept, their significance in determining subject relevancy matters is doubtful.

Either the "simple frequency-based terminology ranking method" or "the averaging of frequency based terminology concurrence ranking method", the approach is only on one directional flow basis (from A/C to B). That is to mean the focus of this approach is on the frequency of occurrence in Group B that contains such similar literature and terminologies also from Group A or Group C. The higher the frequency of the occurrence in Group B the literature or terminologies from Group A or Group C, the stronger their manifestation of their underlying subject connections. If the intended searching approach is from Group A to Group B (or from Group C to Group B), it is based on the assumption that those terminologies of the highest occurrence frequency in Group B have the strongest subject connections with Group A and Group C literature. In reality, however, there is another situation: Relevant terminologies in Group B may be of low-frequency occurrence, which may also concurrently appear in Group A or Group C. In such a case, there is also a strong subject connection among Groups A, B and C, which needs to be factored into the process of knowledge discovery. To do otherwise may result in the potential loss of a significant portion of meaningful subject connections in this tri-angler intellectual relationship of Groups A, B, and C.

3 Bi-directional terminology occurrence-frequency based ranking method

In the knowledge discovery process by way of identifying related terminologies in Group B, the most commonly used ranking or prioritizing approach is based on the frequency of their concurrent presence. From the perspective of the occurrence frequency of certain pertinent terminologies that are related to a given subject, it is the Zipf's Law that provides a theoretical basis for operational guidance, which can be explained as follows: It considers if a word/ terminology usage in a relatively

National Science Library, Chinese Academy of Sciences

long article (exceeding 5 000 words) is ranked in an order of decadency according to its frequency of occurrence and then assign a numerical code to them (i.e. the highest frequency of occurrence of a certain terminology gets assigned to class 1; the next highest frequency of occurrence gets assigned to class 2 and so on). If "f" represents the occurrence frequency, "r" represents ranking order, then, $f_r = c$. (c is a constant)[4]. Based on Zipf's Law, Rune proposed to eliminate both the highest frequency and the lowest frequency terminologies according to a criterion, which made the remaining terminologies to be the most representative for the thematic subject of the given article. That means those terminologies which have both higher frequency and practical significances are the most important elements to be taken into account. In the process of knowledge discovery from non-interactive documents, the researchers integrated the above theory into the concurrence theory. Their task of ranking the subject connections of these terminologies and/or phrases in Group B is based on their high frequency of concurrence, which suggests that there is a strong subject connection of them with Groups A/C. In practice, these terminologies are ranked in a descending order by factoring into consideration the frequency occurrence of either Group B in Groups A (the open model) or Group B in Groups A and C (the closed model) and then setting up a threshold value as a yardstick to admit those terminologies of higher value than the established threshold value in order to be analyzed as Group B and then continue to proceed for knowledge discovery undertakings in Group C eventually.

Aiming at probing the problem of the one-way directional method in dealing with the issue of the occurrence frequency of relevant terminologies, the authors propose a bi-directional ranking method for frequency-based query terminologies as a possible alternative. It means selecting the high-frequency terminologies on the one hand and also at the same time considers their subject relevancy of low-frequency terminologies. To elucidate this scenario, we should examine carefully, first of all, the significance of these low-frequency terminologies and their possible value range as well. In this article, the authors selected both of those relevant terminologies of high-frequency and of low-frequency in Group B and then analyzed the significance and the operational process of this method through an experimentation.

4 An experimentation

4.1 Materials

4.1.1 Test topic

Our test topic comes from the article "A quantitative model for linking two disparate sets of articles in MEDLINE. Bioinformatics, 2007"[5], written by Vetle I. Torvik and Neil R. Smalheiser. The main reason for us to choose this article for testing is that it contains six topics, which we consider them as gold standards

because they are the fruit of the reputable researchers in the medical field themselves.

4.1.2 Arrowsmith system: A knowledge discovery system based on non-interactive documents

The Arrowsmith system, which was created by Swanson and Smalheiser in 1997, is a computer-human interactive system for the purpose of knowledge discovery[6]. It is available on the World-Wide Web. It is designed with capabilities to assist users to extract words and phrases that co-occur in the Group A (start literature) and Group C (target literature) document sets from the closed model. The system is based on a three-way interaction between computer software, a bibliographic database and a human operator[7].

4.1.3 Stop-word lists

Stop-words refer to those words that have no significance but often occur in search texts. Most stop-word lists consist of hundreds of words that have nothing to do with the subject matter of the texts. The experiment mentioned in this article uses MEDLINE stop-word list, which can be obtained for free via its website[8].

4.2 Methods and process

4.2.1 Search strategy

Search each of the six topics mentioned as gold standards in Torvik's article by using the Arrowsmith system[5]. Get query terminologies from Group B collection and have them filtered through by specific semantic groups. The results are shown in Table 1.

4.2.2 Selecting query terminologies of both high-frequency and low-frequency from Group B

After having obtained a desired list of terminologies from Group B through using the Arrowsmith system, then have them screened and then make a selection of the low-frequency terminologies from this Group B. The key problem is how to select the low-frequency occurrence terminologies according to what standards to select. This article initially used a small probability scenario as a determining standard. In the probability theorem, if the probability of the occurrence of an event near to 0, then the event called a small probability event. In general, the values of 0.01 and 0.05 are in common use. That is to say, the probability of the occurrence of an event below 0.01 or 0.05, these events are called small probability events. These two values (0.01 and 0.05) are termed as small probability standards. Therefore, we selected low-frequency occurrence terminologies from Group B according to the small probability standard in our experiment. At the same time, in addition to use

Table 1 Six search strategies and search results by using the Arrowsmith system

Search strategy	Group A	Group C	Group B	Semantic filter and results
1	Retinal detachment[ti]	Aortic aneurysm[ti]	$n=2532$	Procedures; $n=326$
2	Mglur5[ti] OR metabotropic glutamate receptor[ti] OR metabotropic glutamate receptors[ti]	Lewy body[ti] OR Lewy bodies[ti]	$n=1001$	Genes & molecular sequences, and gene & protein names; $n=112$
3	"Magnesium"[mh] AND magnesium[ti] AND ("1900"[pdat]:"1987/12/31"[pdat])	("Migraine disorders"[mh] AND migraine[ti]) AND ("1900"[pdat]:"1987/12/31"[pdat])	$n=1879$	Anatomy; $n=159$
4	Beta-amyloid precursor protein[ti] OR amyloid precursor protein[ti] OR APP[ti] AND ("amyloid"[mh] OR amyloid [tw])	Reelin[All Fields]	$n=1302$	Genes & molecular sequences, and gene & protein names; $n=159$
5	("Nitric oxide"[mh] OR nitric oxide[ti]) AND ("mitochondria"[mh] OR mitochondria[ti] OR mitochondrial[ti])	Psd[ti] OR psd93[ti] OR psd95[ti] OR psds[ti] OR "postsynaptic density" [ti] OR "postsynaptic densities"[ti]	$n=722$	Physiology; $n=59$
6	Calpain[ti] OR "calpain" [mh]	Postsynaptic[All Fields] AND density[All Fields]	$n=3496$	Genes & molecular sequences, and gene & protein names; $n=379$

small probability events as the selecting standard in this research, we also selected low-frequency occurrence terminologies to compare with those of high-frequency occurrence ones. It is believed that such a method of selecting the subject-relevant terminologies would be more meaningful in actuality.

Table 2 shows the amount of the selected terminologies which has a probability of selection $\geqslant 0.9$ and $\leqslant 0.05$ and their ratio of percentage in Group B. Generally, the average percentage of terminologies in high-frequency occurrence is approximately 20% and the percentage of those in low-frequency occurrence is 40% in Group B. Thus it can be seen that the number amount of selected terminologies in this field range was large enough for targeted research purpose on the one hand and still within the range of manageability in terms of statistical analysis on the other hand. After having completed the selection of terminologies in low-frequency occurrence in Group B, we filtered out those terminologies that have a connotation too broad to be of any practical significance. As a result, we extracted 213 terminologies of high-frequency occurrence and 270 terminologies of low-frequency occurrence from Group B that had practical implications to our research. As for the 483 terminologies so retrieved, we devised six search strategies stemming from the various combinations of the retrieved terminologies of both high and low frequency

occurrence from segments in Group B. Specifically, the composition of each of these six search strategies are as follows: 1) A combination of 4 high-frequency occurrence terminologies and 64 low-frequency occurrence terminologies; 2) 31 high-frequency occurrence terminologies and 40 low-frequency occurrence terminologies; 3) 34 high-frequency occurrence terminologies and 47 low-frequency occurrence terminologies; 4) 33 high-frequency occurrence terminologies and 29 low-frequency occurrence terminologies; 5) 5 high-frequency occurrence terminologies and 27 low-frequency occurrence terminologies; and 6) 36 high-frequency occurrence terminologies and 63 low-frequency occurrence terminologies.

Table 2 Probability $\geqslant$0.9 and probability $\leqslant$0.05 of the distribution of the retrieved terminologies in Group B

Search strategy	Group B	High-frequency from Group B	Ratio (%)	Low-frequency from Group B	Ratio (%)
1	326	80	24.54	126	38.65
2	112	31	27.68	43	3839
3	159	41	25.79	64	40.25
4	159	34	21.38	62	38.99
5	59	5	8.47	34	57.63
6	379	52	13.72	151	39.84

4.2.3 Selecting terminologies of co-concurrent existence in Groups A and/or C against those of high and low frequency occurrence in Group B

We obtained articles from Groups A and C, by using the retrieved search terminologies of both high and of low frequency occurrence from Group B in conjunction with the Arrowsmith system. The occurrence frequency of relevant terminologies in Group A and/or C with those of Group B was tallied. After comparing the differences of the occurrence frequency of relevant terminologies in Groups B with those in Group A or C, we took those documents, which had an occurrence frequency of terminologies more than 5 or 10 times for an analysis about their medical significance. The reason we chose 5 times as the threshold value standard was that when the occurrence frequency of terminologies is more than 5 times, it would be considered as a relatively high frequency of occurrence in the subject-relevant fields. In addition, the occurrence frequency of terminologies more than 5 times is of the right percentage of all selected documents contained in Groups A, B, and C.

4.3 Results and analyses

By a statistical analysis of those articles retrieved, the results can be shown in Tables 3 and 4. There were no significant differences between the occurrence frequency of terminologies of Groups A and C in the high or low frequency occurrence of Group B. Generally speaking, the occurrence frequency or relevant

National Science Library, Chinese Academy of Sciences

terminology in Group A or C is higher than those in low-frequency component of Group B.

Furthermore, our analysis showed that those articles retrieved from Groups A and/or C had a higher occurrence frequency (i.e., the frequency is more than 5 and 10 times). Articles of occurrence frequency in Groups A and C more than 5 times accounted for nearly 29.78% of the total amount in the high-frequency occurrence sector of Group B. Articles in Groups A and C had an occurrence frequency more than 10 times in the high-frequency occurrence sector of Group B accounted for nearly 12.67% of the total amount of articles.

Articles in groups A and B which had an occurrence frequency more than 5 times in the low-frequency sector of Group B accounted for about 34.64% of the total amount of articles. Articles in Groups of A and/or C, which had an occurrence frequency more than 10 times in the low-frequency sector of Group B accounted for about 12.69% of the total amount of articles. In our further research, we analyzed 14 terminologies, contained in Groups A or C that had subject connections with those of high frequency occurrence sector of Group B. We found that there were six relevant terminologies that exceeded the average number of such terminologies contained in Groups A or C that had a rate of occurrence frequency more than 5 times. However, in the low-frequency of relevant terminology occurrence sector of Group B, the occurrence frequency of those documents in Groups A and C more than 5 times, which had six more terminologies that surpassed the average amount of the total number of subject relevant terminologies. Furthermore, among those subject relevant terminologies in the low-frequency sector of Group B, there were some relevant query terminologies extracted from Groups A or C, which had a conspicuous higher percentage of occurrence frequency. They include such query terminologies as mitochondria, retinal detachment, aortic aneurysm, metabotropic glutamate receptors, migraine disorders, postsynaptic and density. These terminologies accounted for half of the total documents in Groups A or C.

From the above analysis, we can see the situation that documents in the low frequency sector of Group B does contain documents of Groups A and/or C in high frequency of occurrence. The percentage of such a situation of occurrence is quite large actually. If this group of articles is overlooked or arbitrarily deleted in the process of knowledge discovery from non-interactive resources, it is quite possible that we may lose a significant portion of subject connections that is beneath the surface of researchers' visibility and/or imagination.

We only quantitatively analyzed the significance of documents occurrence in low-frequency in Group B as shown in Tables 3 and 4. For a further explanation of the significance of documents in low-frequency occurrence component in Group B, we took a further step to have the medical significance of those articles from Groups A and C that appeared in the low frequency occurrence component of Group B analyzed. The result is shown in Table 5. It shows that in the low frequency

Table 3 Group A/C documents occurrence frequency distribution in the component of high-frequency occurrence in Group B

Group A/Group C	Articles	Total frequency	Frequency	Articles of frequency ⩾5		Articles of frequency⩾10	
				Number	Percentage (%)	Number	Percentage (%)
Mitochondrial	156	493	3.16	43	27.56	6	3.85
Nitric oxide	156	298	1.91	13	8.33	3	1.92
Psd	5	42	8.4	4	80.00	1	20.00
Amyloid precursor protein	152	302	1.99	122	80.26	46	30.26
Reelin	72	302	4.19	27	37.50	9	12.50
Retinal detachment	134	583	4.35	55	41.04	16	11.94
Aortic aneurysm	135	481	3.56	32	23.71	12	8.89
Metabotropic glutamate receptors	248	546	2.2	43	17.34	12	4.84
Lewy body	388	1,242	3.2	170	43.81	42	10.82
Magnesium	215	357	1.66	15	6.98	6	2.79
Migraine disorders	119	357	3	18	15.13	7	5.88
Calpain	225	767	3.41	42	18.67	26	11.56
Postsynaptic	388	2,064	5.32	114	29.38	98	25.26
Density	290	1,253	4.32	101	34.83	56	19.31
Total	2,683	9,087	3.39	799	29.78	340	12.67

Table 4 Group A/C document occurrence frequency distribution in the component of low-frequency occurrence in Group B

Group A/Group C	Articles	Total frequency	Frequency	Articles of frequency ⩾5		Articles of frequency⩾10	
				Number	Percentage (%)	Number	Percentage (%)
Mitochondrial	193	733	3.8	67	34.72	20	10.36
Nitric oxide	190	367	1.93	19	10.00	11	5.79
Psd	92	511	5.56	55	59.78	12	13.04
Amyloid precursor protein	394	886	2.25	135	75.00	55	30.56
Reelin	102	359	3.52	12	11.76	10	9.80
Retinal detachment	187	851	4.55	87	46.52	28	14.97
Aortic aneurysm	192	684	3.56	82	42.71	28	14.58
Metabotropic glutamate receptors	212	954	4.5	79	37.26	41	19.34
Lewy body	178	498	2.8	50	28.09	18	10.11
Magnesium	758	1,514	2	80	10.55	28	3.69
Migraine disorders	90	286	3.18	25	27.78	7	7.78
Calpain	326	1,339	4.11	46	14.00	12	3.68
Postsynaptic	457	2,824	6.18	324	71.00	98	21.44
Density	278	1,857	6.68	203	73.03	95	34.17
Total	3,649	13,663	3.74	1,264	34.64	463	12.69

National Science Library, Chinese Academy of Sciences

occurrence component of Group B, if the occurrence frequency of articles from Groups A and C was larger than five times, 72.39% of those articles are medically meaningful. By the same token, if such documents occurrence frequency is over ten

times, then 83.15% of them have medical significance, of which 5 document clusters even have reached a perfect score of 100%. From this research finding, we can see that there is not only the existence of relevant articles from the high-frequency occurrence components of Groups A or C in the low frequency occurrence component of Group B, but also in a high proportion. In addition, it is indicated that from a perspective of the thematic contents of those documents, approximately 80% of Groups A-B connections or Groups C-B connections in the low frequency occurrence component of Group B has significance in medical science.

Table 5 Significant articles of occurrence frequency from Groups A/C $\geqslant$5 times and also their frequency $\geqslant$10 times in low-frequency component of Group B

Group A/Group C	Articles of frequency $\geqslant$5	Significant article numbers	Ratio (%)	Articles of frequency $\geqslant$10	Significant article numbers	Ratio (%)
Mitochondrial	67	43	64.55	20	14	79.00
Nitric oxide	19	14	73.68	11	11	100.00
Psd	55	35	64.29	12	9	75.00
Amyloid precursor protein	135	91	67.44	55	49	89.09
Reelin	12	9	75.00	10	10	100.00
Retinal detachment	87	70	80.00	28	23	82.14
Aortic aneurysm	82	51	62.20	28	28	100.00
Metabotropic glutamate receptors	79	63	79.75	41	29	70.73
Lewy body	50	36	72.00	18	15	83.33
Magnesium	80	61	76.25	28	28	100.00
Migraine disorders	25	19	76.00	7	7	100.00
Calpain	46	30	65.22	12	12	100.00
Postsynaptic	324	243	75.00	98	74	75.51
Density	203	150	73.89	95	76	80.00
Total	1,264	915	72.39	463	385	83.15

5 Conclusion

On the basis of the above analysis, there are documents from either Group A or Group C, which have a high frequency of occurrence in the low-frequency document component of Group B. In addition, with the exception of only a few isolated situations, those articles of high frequency occurrence in Groups A or C took a respectable percentage. They averaged about 30% of the total amount of articles. Some of the high frequency occurrence articles from Group A reached a percentage even over 50%. After our analyzing the significance of subject connections, a few articles, which had a high frequency of occurrence in groups A and/or C, were filtered out as so to reduce their percentage of presence. However, there were relatively more articles of true medical significance among those articles which had a high frequency of occurrence; the percentage of such articles exceeds 70%. When the frequency of occurrence of meaningful terminologies is raised to 10 times, the

percentage of the more significant articles also increase proportionately. Some articles about a certain subject may pose significant connections to all articles on that subject such as, magnesium, aortic aneurysm, nitric oxide, calpain, reelin, etc. From above research findings, we can see that there are terminologies of high frequency occurrence in Group A and/or C that also concurrently exist in the low frequency occurrence component of Group B. Furthermore, the probability of having significant subject correlations among these articles in the high-frequency occurrence components of A and/or C is higher. Consequently, in the process of seeking knowledge discovery from non-interactive documents and in selecting those terminologies of having concurrent appearance in low frequency occurrence component of Group B, we should also consider the desirability and significance of extracting some appropriate terminologies and documents from the low-frequency component of Group B. Such a procedure will improve the quality of Group B and increase the efficiency of knowledge discovery consequently. Therefore, in the process of ranking the relevancy of terminologies and documents in Group B, we should use the bi-directional frequency-based ranking method for knowledge discovery purposes. As for how we should determine the specific values of certain terminologies and documents in low frequency occurrence components and whether such of our value assessment undertakings based on our different knowledge discovery models will create more problems, etc., are some of the unresolved important issues, which need to be further studied in depth in the near future.

References

1 Swanson, D. R. Undiscovered public knowledge. Library Quarterly, 1986, 56:103–118.
2 Rong, Y. H. & Liang, Z. P. Literature-based knowledge. Journal of the China Society for Scientific and Technical Information (in Chinese), 2002, 21(4):386–390.
3 Ma, M. & Wu, Y.S. Methodological enlightenment and significance of Don R. Swanson's achievements in information science. Journal of the China Society for Scientific and Technical Information (in Chinese), 2003, 22(3):259–266.
4 Qiu, J. P. Informetrics (in Chinese). Wuhan: Wuhan University Press, 2007.
5 Torvik, V. I. & Smalheiser, N. R. A quantitative model for linking two disparate sets of articles in Medline. Bioinformatics, 2007, 23(13):1658–1665.
6 Swanson, D. R. & Smalheiser, N. R. An interactive system for finding complementary literature: A stimulus to scientific discovery. Artificial Intelligence, 1997, 91:183–203.
7 Arrowsmith. Retrieved on December 1, 2009, from http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html.
8 Medline Stopwords. Retrieved on December 1, 2009, from http://kiwi.uchicago.edu/stopwords_pubmed.

(*Copy editor*: *Ms. Jing CAO; Language revison: Prof. Charles C. YEN*)

National Science Library,
Chinese Academy of
Sciences