Progress Reports

Research on the development of thesaurus in China[①]

GAO Wenfei[1]* & ZHAO Xinli[2]

[1] Beijing Information Resource Management Center, Beijing 100101, China
[2] China Science and Technology Exchange Center, Beijing 100045, China

Translator: Editorial Office of *CJLIS*

Received Dec. 22, 2008
Revised Feb. 10, 2009
Accepted Feb. 16, 2009
Translated with a permission from *Library and Information Service* (in Chinese), 2008, 52(9):66

Abstract Based on the statistics of 130 thesauri having been published in China so far, this article analyzes the development of thesauri in China from the origin, publication year, academic disciplines and quantity of entries collected, and re-defines the development stages. In addition, by collecting the 1,000 relevant research papers, the article also analyzes the theoretical studies from the aspects of the quantity of papers and research subjects in order to give a clear picture of the development and features of the researches on thesaurus in China.

Keywords Thesaurus, Ontology, Development stage

In recent years, it has become a common concern of study to introduce ontology into the information retrieval system as a new tool of information organization. And, thesaurus and ontology share many similarities in terms of the function and theory basis. If the existing thesauri can be transformed into corresponding ontologies, the ontology building will be greatly improved[1]. As such, it is necessary for us to have a deep understanding of the development of thesaurus in China, while studying the theory and application of ontology. Based on the statistics of 130 thesauri having been published in China so far, this article analyzes the development of existing thesauri in China from the origin, publication year, distribution of academic disciplines and quantity of entries collected, and re-defines the development stages. In addition, by collecting the 1,000 relevant research papers, the article also analyzes the theoretical studies from the aspects of the quantity of papers and research subjects in order to give a clear picture of the development and features of the researches on thesaurus in China.

1 Publication year

The publication year of 130 thesauri this article collecting was shown in Table 1 and Fig. 1.

CJLIS
Vol. 1 No. 3/4, 2008
pp 128–138
National Science Library, Chinese Academy of Sciences

[①] This work is supported by the National Nature Science Foundation of China (Grant No. 70573103).

* Correspondence should be addressed to Gao Wenfei (E-mail: gaowf@beijingit.gov.cn).

Table 1 Distribution of publication year of thesauri in China

Year	Quantity of published thesauri	Percentage (%)	Average / Year
1956–1965	1	0.77	0.1
1966–1975	1	0.77	0.1
1976–1985	22	16.92	2.2
1986–1995	82	63.08	8.2
1996–2005	24	18.46	2.4

Note: As the first edition of *Thesaurus for Aerospace Sciences (Hangkong Keji Ziliao Zhuti Biao)* published in 1964 differs from the revised edition published in 1971 in both the feature and infrastructure, they are usually considered as two separate thesauri by statistics. However, to avoid misunderstanding, the other thesauri are measured only by their first edition in terms of statistics.

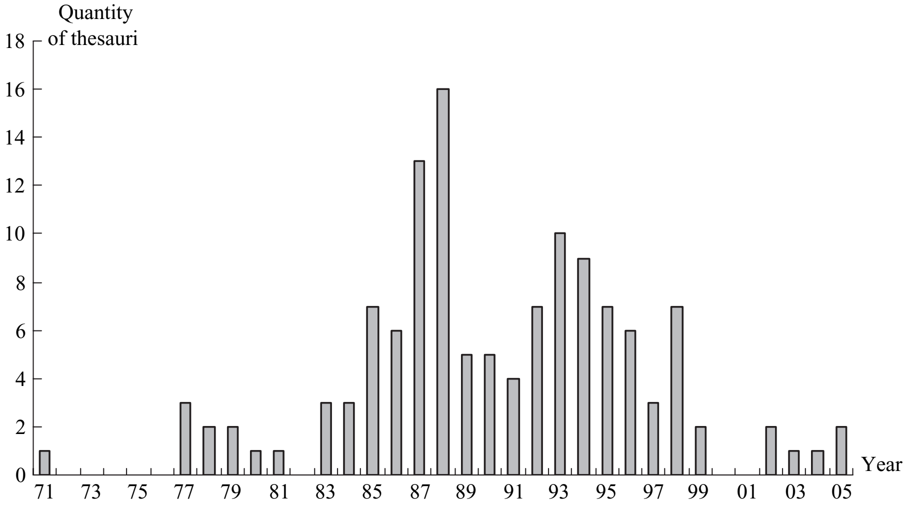

Fig. 1 The change of the quantity of thesauri published in China.

Figure 2 reflects the development stages of thesauri in China, from which it can be seen that the period from 1956 to 1965 is the testing stage. Although the *Guideline for Chinese Subject Headings (Zhongwen Tushu Biaoti Fa)* edited by Cheng Changyuan was published in 1950, there were very few practices in China. However, since 1956, the booming of the scientific and cultural undertakings had promoted the development of thesauri in China. *The Subject Headings of Aerospace Sciences (Hangkong Keji Ziliao Zhuti Biao)*, the first subject heading list in China, was finished in May 1964, which, consisting of eight separate volumes, was translated and adapted on the basis of the *Thesaurus of Armed Services Technical Information Agency* (ASTIA). *The Chinese Pinyin & Character Index of Subject Headings (Zhutici Hanyu Pinyin Zishun Suoyin)* and *the Chinese Pinyin & Character Index of sub-headings (Zi biaotici Hanyu Pinyin Zishun Biao)* were published in January and August 1965 respectively[2]. The former Information Institute of Academy of Aerospace Sciences (Hangkong Yanjiuyuan Qingbao Suo) took the

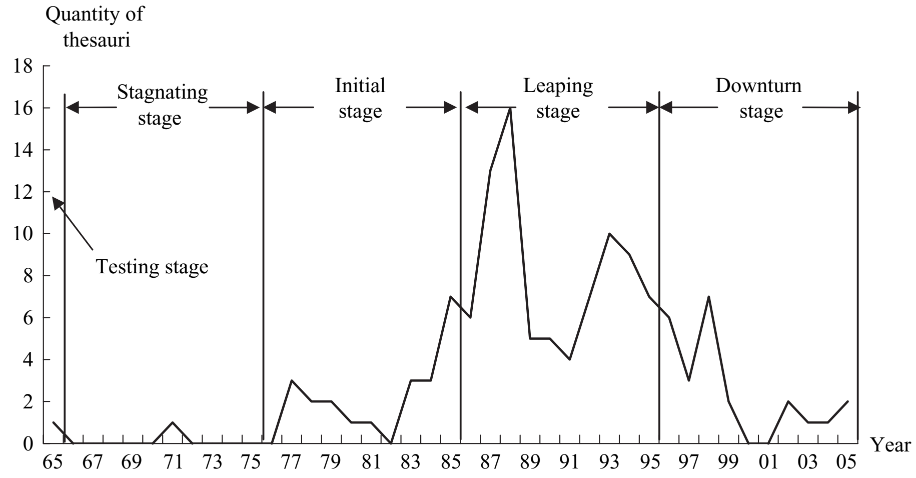

Fig. 2 The developing stages of thesaurus in China.

lead in using the Thesaurus to index its whole collection and further promoting it within the information institutes or offices in the aerospace science and technology industry.

The development had stagnated from 1966 to 1975. The publication of *Thesaurus of Aerospace Sciences (Hangkong Keji Ziliao Zhuti Biao)* and its application in the document indexing in some institutes should have greatly pushed the development of the thesaurus. However, due to the Cultural Revolution, the development was almost at a standstill during the period. It was only in 1971 that the Information Institute of Academy of Aerospace Sciences (Hangkong Yanjiuyuan Qingbao Suo) revised the *Subject Headings of Aerospace Sciences (Hangkong Keji Ziliao Zhuti Biao)* and transformed it from subject headings to the thesaurus. In 1972, the Information Institute of the Commission of Science, Technology and Industry for National Defense (*Guofang Kewei Qingbao Suo*) started to compile the *Subject Headings of Science and Technology for National Defense* (*Guofang Kexue Jishu Zhuti Cidian*).

The period from 1976 to 1985 marked the initial stage of the development, during which 22 thesauri were compiled, accounting for 16.92 percent of all the thesauri collected. The *Chinese Thesaurus (Hanyu Zhutici Biao)* compiled and published by the Institute of Scientific and Technological Information of China (ISTIC) in 1979 became an important milestone and gave a great impetus to the development of thesauri in China both theoretically and practically[3]. During this period, many thesauri were published on the basis of that, such as *Chinese Thesaurus for Chemical Engineering (Huagong Hanyu Zhutici Biao)*, *Chinese Thesaurus for Geology (Dizhixue Hanyu Xuci Biao)* and *Chinese Petroleum Thesaurus (Shiyou Gongye Hanyu Zhutici Biao)*, etc.

The period from 1986 to 1995 was the leapfrog stage, during which 82 thesauri were compiled, accounting for 63.08 percent of all the thesauri collected. In particular, a rapid development was witnessed in the late 1980s, as evidenced by 13 thesauri compiled in 1987 and 16 thesauri finished in 1988. In 1990, the *Thesaurus of Military (Junyong Zhutici)* was published, and later a series of military thesauri were successively compiled. Among the total 22 military thesauri collected, 15 thesauri were compiled in this period. It was in this very period that China started emphasizing the standardization of the thesaurus and formulated the *Guidelines for Establishment and Development of Chinese Thesauri* (GB13190-1991), *Documentation — Guidelines for Establishment and Development of Multilingual Thesauri* (GB15471-1994) and *Guidelines for Establishment and Development of Military Thesauri* (*Junyong Zhutici Biao Bianzhi Guize*) (GJB1776-93).

With the researches on classification-subject integration and relevant practices developing, a batch of classified thesauri were compiled during this period, such as *China Classified thesauri* (*Zhongguo Fenlei Zhutici Biao*), *Agricultural Science Thesaurus (Nongye Kexue Xuci Biao)* and *Social Science Thesaurus* (*Shehui Kexue Jiansuci Biao*). Classification-subject integration is considered as one of the major developments of the information retrieval. Up to now, China has compiled nearly 20 classified thesauri, all of which are facet thesauri, except *China Classified Thesaurus* (*Zhongguo Fenlei Zhutici Biao*). The facet and classification relations in facet thesaurus are demonstrated in a more systematic, comprehensive and defined way than in that of the category list and hierarchy list, so the facet thesauri show better practicality[4]. Therefore, it is expected that transforming the existing thesaurus into the facet thesaurus should be one of the future development trends.

From 1996 to 2005, the compilation of China's thesauri entered into a downturn stage, during which only 24 thesauri were compiled, accounting for 18.46 percent of the total. During that period, China's thesauri developed from the compilation of new thesauri to the revision of the existing thesauri, as evidenced by the *Guidelines for Establishment and Development of Military Thesaurus* (*Junyong Zhutici Biao Bianzhi Guize*) (GJB1776A-99) published in 1999 to replace the 1993 edition, and the *Guidelines for the Establishment and Development of Electronic Military Thesaurus* (GJB 5098-2004) issued in 2004 to replace the 1999 edition. With the rapid development of E-governance in China in the 21st century, the *E-Governance Thesaurus* (*Zonghe Dianzi Zhengwu Zhutici Biao*) and the *Guidelines for the Establishment and Development of E-Governance Thesaurus* (*Dianzi Zhengwu Zhutici Biao Bianzhi Guize*) (GB/T 19486-2004) were published respectively in 2005.

With the constant changes in the production of information resources, the information environment, the user community and the retrieval requirement, there are many drawbacks in thesauri, such as simple semantic relations, poor interoperability and representation, and non-formalization problems. Under such circumstance, the researches on thesauri during this period focused on the application on Internet and seeking for new forms of thesauri. At present, the development of ontology derived from thesaurus has become a hot issue, and more than 10 kinds of thesauri have been converted into ontology via various methods in foreign countries. However, there has been no applicable ontology available so far in China, although some positive attempts have been made in this regards. The facet thesauri with more clear semantic relationship and more flexible structure can be converted into ontology more easily, compared with the traditional thesauri. Therefore, it should be an effective method worth trying to build ontology by first transforming the existing thesauri into the facet thesauri, and then building ontologies based on them.

2 Distribution of academic disciplines

The 130 thesauri collected can be classified in terms of academic disciplines, as shown in Fig. 3. From the perspective of the academic disciplines of the thesauri, there are 31 thesauri in social sciences, accounting for 23.85 percent; 88 thesauri in natural sciences, accounting for 67.69 percent; and 11 thesauri covering both social sciences and natural sciences, accounting for 8.46 percent. There are 11 comprehensive thesauri in social sciences, accounting for 35.48 percent of the thesauri in social sciences and 8.46 percent of the total thesauri, and 10 comprehensive thesauri in natural sciences, accounting for 11.36 percent of the thesauri in natural sciences and 7.69 percent of the total thesauri. In total, there are 32 comprehensive thesauri, accounting for 24.62 percent of the total. That has shown that China has not only compiled a large number of academic thesauri focusing a specific field, but also is heading towards the compilation of comprehensive thesauri.

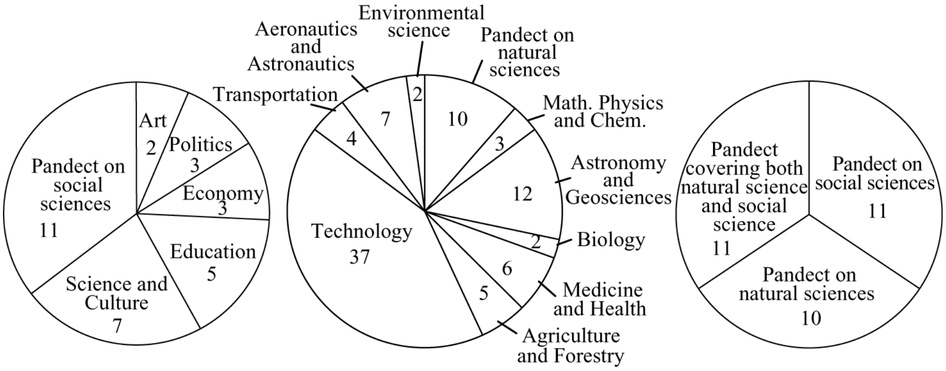

Fig. 3 Distributions of academic disciplines of thesauri in China.

By comparing the quantity of thesauri in social sciences and in natural sciences, it is found that there are 88 thesauri in natural sciences and only 31 thesauri in social sciences. And among the thesauri in natural sciences, there are 37 thesauri in the field of industrial technology. This indicates the unbalanced development of thesaurus in China. The compilation of thesauri focuses mainly on natural sciences, especially in the field of the industrial technology, while the work in social sciences is yet to be developed. In addition, it is worth mentioning that among the 130 collected thesauri,

National Science Library, Chinese Academy of Sciences

there are 22 military thesauri, showing a sustainable development of China's military thesaurus.

3 Quantity of entries collected

The thesaurus can be divided into three types according to the quantity of entries collected. Large thesaurus has more than 10,000 entries collected, medium-sized thesaurus has entries collected between 1,000 and 10,000, and small-sized thesaurus has less than 1,000 entries collected[5]. Among the 130 thesauri collected, 89 thesauri with detailed statistics are shown in Table 2.

Table 2 Thesauri in terms of the collected entries

	Large thesaurus	Medium-sized thesaurus	Small-sized thesaurus
Quantity	32	49	8
Percentage (%)	35.95	55.06	8.99

It can be seen from Table 2 that among 89 thesauri, the medium-sized thesauri are the most, accounting for 55.06 percent followed by the large thesauri, accounting for 35.95 percent, and the small-sized thesauri took up the least, accounting for 8.99 percent only. Among the 32 large thesauri, each of 14 thesauri has more than 20,000 entries collected, among which, *Chinese Thesaurus* (*Hanyu Zhutici Biao*) has collected 108,568 entries, and *Military Thesaurus(Junyong Zhutici Biao)* has collected 52,500 entries. However, there are so few small-sized thesauri, most of which were compiled by government departments or institutions for their internal use only and not released to the public.

As shown in Table 3, the annual number of the three types of thesauri is on the rise, in which medium-sized thesaurus has the fastest growth. From the publication year of large thesauri, more than one half of the 32 large thesauri collected were compiled during the period from 1986 to 1995. Of all the 23 thesauri compiled during the period from 1996 to 2005, 7 thesauri were large thesauri. It can be seen clearly that the medium-sized thesaurus has always occupied a dominant position in different development stages. However, with the development of the Internet, the increase in the amount of literature and the application of the computer-aided technology, the large thesaurus is prevailing worldwide, in which China has yet a long way to go.

4 Literature analysis on thesauri

With the development of thesaurus, a quantity of research papers on thesaurus has been published by China's library and information communities. The author has thus downloaded and sorted out 1,000 research papers on thesaurus researches recorded from 1980 to 2005 from the Chinese Journal Full-text Database (CJFD) in China National Knowledge Infrastructure (CNKI) and the National Index to Chinese Newspaper & Periodicals, then analyzed the researches on thesaurus in that period from two aspects: the quantity of literature and research subjects.

Table 3 Thesauri with quantity of collected entries and distributions of publication year

Year	Large thesaurus			Medium-sized thesaurus			Small-sized thesaurus		
	Quantity of thesaurus	Percentage in this type (%)	Annual average quantity	Quantity of thesaurus	Percentage in this type (%)	Average / Year	Quantity of thesaurus	Percentage in this type (%)	Average / Year
1966–1975	0	0	0	1	2.04	0.1	0	0	0
1976–1985	8	25	0.8	5	10.20	0.5	1	12.5	0.1
1986–1995	17	53.13	1.7	32	65.31	3.2	2	25	0.2
1996–2005	7	21.87	0.7	11	22.45	1.1	5	62.5	0.5

Table 4 Statistics of quantity of research literatures on thesaurus

1976–1985	Year	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	Total	Annual average quantity	Percentage
	Quantity	–	–	–	–	9	16	15	26	51	29	146	24.33	14.6%
1986–1995	Year	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	Total	Annual average	Percentage
	Quantity	34	60	38	51	30	26	20	43	56	64	422	42.2	42.2%
1995–2005	Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	Total	Annual average	Percentage
	Quantity	56	45	33	39	24	32	44	47	62	50	432	43.2	43.2%

4.1 Statistical analysis of the quantity of literature

Statistics of the 1,000 articles collected in terms of publication year is shown in Table 4.

Prior to the 1980s, the researches focused mainly on the subject indexing and catalogs. There was almost no research on thesaurus. The publication of *Chinese Thesaurus* in 1979 not only gave rise to the compilation of a large batch of thesauri, but also promoted the development of the pertinent theoretical researches. Since 1980, a large number of research articles on thesauri have been successively published.

It can be seen from Fig. 4 that the publication year of research literature on China's thesauri coincides with that of the thesauri. During the period from 1980 to 2005, four large-scale academic meetings on information retrieval were held successively, namely, the National Seminar on the Development of Chinese Thesauri held in December 1988, the 2nd National Seminar on the Development of Classification and Thesaurus held in July 1996, the 3rd National Seminar on the Development of Information Retrieval Language in June 1999 and the 4th National Seminar on the Development of Information Retrieval Language in June 2005. As shown in Fig. 4, the number of published articles on thesaurus reached its peak shortly before or after the above-mentioned meetings.

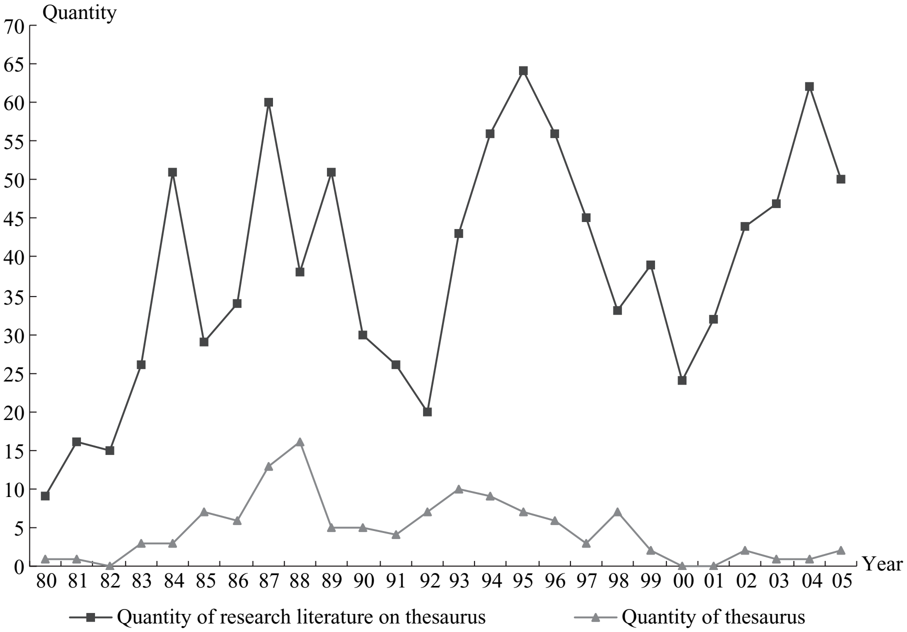

Fig. 4 The Comparison between quantity of research literature on thesaurus and quantity of compiled thesaurus.

4.2 Analysis of the subjects of literature

As we divide all the articles into nine subjects and make further statistical analysis (see Table 5), it can be seen that among the 1,000 articles published from 1980 to 2005, there are 270 articles on review, introduction and compilation of thesauri, followed by 221 articles on classification-subject integration, and 218 articles on theoretical research of thesaurus. The number of the articles on standardization & compatibility and computerization is comparatively small, which are 89 and 60 respectively. Because of China's huge quantity of thesauri, the number of research articles on review, introduction and compilation of thesauri is comparatively large,

Table 5 Statistics of research subjects of the literatures

Research Subject		Year			Total
		1980–1985	1986–1995	1995–2005	
Summary of development	Quantity	2	32	15	49
	Average / Year	0.3	1.5	3.2	1.9
	Percentage	1.4%	7.6%	3.5%	4.9%
Theoretical research	Quantity	74	81	63	218
	Average / Year	12.3	8.1	6.3	8.4
	Percentage	50.8%	19.2%	14.6%	21.8%
Standardization & compatibility	Quantity	6	29	25	60
	Average / Year	1	2.9	2.5	2.3
	Percentage	4.1%	6.9%	5.8%	6%
Computerization	Quantity	11	30	48	89
	Average / Year	1.8	3	4.8	3.4
	Percentage	7.5%	7.1%	11.1%	8.9%
Classification-subject integration	Quantity	6	93	122	221
	Average / Year	1	9.3	12.2	8.5
	Percentage	4.1%	22.1%	28.2%	22.1%
Review, introduction and compilation of thesauri	Quantity	47	157	66	270
	Average / Year	7.8	15.7	6.6	10.4
	Percentage	32.2%	37.2%	15.3%	27%
Application and development in network-based environment	Quantity	0	0	53	53
	Average / Year	0	0	5.3	2.1
	Percentage	0	0	12.3%	5.3%
Integration with natural language	Quantity	0	0	17	17
	Average / Year	0	0	1.7	0.7
	Percentage	0	0	3.9%	1.7%
Related with ontology	Quantity	0	0	23	23
	Average / Year	0	0	2.3	0.9
	Percentage	0	0	5.3%	2.3%
Total	Quantity	146	422	432	1000
	Average / Year	24.3	42.2	43.2	38.46
	Percentage	14.6%	42.2%	43.2%	100%

of which those on *Chinese Thesaurus* (*Hanyu Zhutici Biao*) forms a major proportion. Classification-subject integration is a key subject of China's thesaurus research, the literature of which has been growing especially since 1986 when the relevant articles increased sharply due to the publication of *China Classified Thesaurus* (*Zhongguo Fenlei Zhutici Biao*). The articles published on this topic during the decade from 1995 to 2005 reached to 122, accounting for 28.2 percent of the total articles published in that decade. The researches on standardization & compatibility started fairly early in China, but unfortunately the relevant articles are comparatively few. Therefore, we should pay more attention to those researches in future. Despite the few of the articles on computerization, relevant articles are expected to grow with the constant development of the computer technology.

In addition, it can be also seen from Table 5 that, since the late 1990s, new development trends emerged. With the development of information technology and network technology, especially those applications on information retrieval, thesauri have been transferred from print edition to electronic format, then to online version. Hence, the researches on the application and development of thesaurus in the network-based environment have become a main topic. During the decade from 1995 to 2005, there are 53 articles on that topic, accounting for 12.3 percent of the total published. In the network-based environment, the information technology contributes greatly to the development of the natural-language retrieval. Since 1999, the research articles on integrating the subject headings with natural language have been successively published. It is expected to be a key research topic on the theories, methods and technologies of the integration of subject headings and natural language will become an important part of the researches on retrieval in subject headings and natural language. Furthermore, it can be also seen from the distribution of academic disciplines that the integration of thesauri and ontologies becomes another important development trend. Since 2003, many researchers have written papers to explore the relations between thesaurus and ontology, and the ontology building based on the thesauri. It has become a current hot issue to embed ontology into the retrieval system, and that will be a future research focus.

In conclusion, with the rapid development of theories and technologies of the thesauri, and with the constant updating of man's knowledge, all kinds of new technologies are expected to be applied to revise and maintain the thesauri continuously, and further innovations are expected to be made in the networking environment. We should prepare to keep up with the new emerging developments, and pay more attention to the study of new theories, new types of information organization such as ontology, interoperability and standardization of thesauri and application in networking environment, etc. which are all key research topics in the future.

References

1 Wang, S. F. An introduction to the integration of the controlled vocabulary and ontology. Journal of Academic Libraries (in Chinese), 2005(1):74–78.

2 Qiu, Z. B. Thesaurus for aerospace sciences and its application (in Chinese). Beijing: China Aerospace Information Center, 1999.

3 Dai, W. M., Zhao, J. B., & Wang, D. B. The information linguistics research faced to 21 Century (in Chinese). Beijing: Beijing Library Publishing House, 2000.

4 Zhang, Q. Y. Faceted reform of thesaurus. Library Tribune (in Chinese), 1997(6):30–31.

5 Hou, H. Q., & Xu, J. The survey and developing trend of thesaurus in foreign countries. Information Science (in Chinese), 1989, 8(5):378–386.