A Summary of Thesaurus Research Based on Co-Word Analysis

Jinhua Yao[1,a,*]

[1]*College of Economics and Management, Nanjing University of Aeronautics and Astronautics,Nanjing, China*
a. yjh907@163.com
**corresponding author: Jinhua Yao*

Keywords: Thesaurus, co-word analysis, clustering, reseacrh status introduction.

Abstract: At present, with the rapid development of big data, cloud computing, and the semantic web, in order to meet the needs of real-time processing of big data and information, it is necessary to strengthen the research work of knowledge organization systems. Based on the bibliometric and co-word clustering analysis method, this paper selects the subject of library information and digital library under the China national knowledge infrastructure database, and uses the periodicals collected in 2010-2019 as the research object. Through the visualization of knowledge graphs, the analysis of domestic descriptors in the past ten years research status in the field of thesaurus. The analysis results show a slight downward trend in thesaurus research. However, it has not been abandoned as an effective knowledge organization tool, and its development has shown a trend of networking, diversification and integration.

1. Introduction

With the development of the big data, traditional knowledge organization and knowledge sharing methods are becoming more and more difficult to meet people's increasingly diverse needs. In recent years, with the construction of knowledge bases, the emergence of "intelligent search engines", and the development of ontology, etc., all indicate that the Internet will enter the "semantic age". As a kind of information retrieval language, the thesaurus mainly acts on the organization, retrieval and utilization of information. In the network environment, the thesaurus highlights the huge research value of the thesaurus in the fields of intelligent retrieval, machine translation, and semantic reasoning. In order to explore the frontiers and trends of the thesaurus research, this article uses the co-occurrence analysis method to carry out bibliometric analysis, combs and analyzes the research topics of the thesaurus in the domestic library and information field, and more intuitively and concretely presents the research status of the domestic thesaurus.

The thesaurus, also known as the index dictionary, is a retrieval tool for indexing topics in document and information retrieval; compared to natural language, it is a standardized, organized and subject content a collection of defined terms. The term "thesaurus" was first applied to the field of information retrieval by Peter Lunch of International Business Machines Corporation (IBM) in 1957[1]. The first thesaurus came out in 1959 and was compiled by DuPont[2]. In the early 1960s, computer technology was introduced into the field of library and

information, which strongly promoted the development of the thesaurus. After that, hundreds of thesaurus were published, and the automation of thesaurus compilation is also increasing ,which promotes the development of computer innformation retrieval technology. In the 1970s, China gradually began related research on the thesaurus and compiled a number of comprehensive and professional thesaurus, which better met the organization and retrieval needs of various types of units, and improved the efficiency of information utilization.

Starting with several key concepts and things in the thesaurus research, Yu Fengmin[3] proposed a research evolution path with the purpose of knowledge organization and the application technology of the thesaurus. Zhang Zhongqiu[4] conducted a survey of journal articles related to the classification table and the thesaurus in china from 2005 to 2011, and analyzed from six aspects including the related research on the ontology, the sorting promoted by KOS, and the research on the development of the subject indexing and retrieval method Zeng Jicheng[5]used the methods of bibliometrics and information visualization to sort out the research on the thesaurus in China in recent years, and using the Citespace theme cluster analysis the key fields of automatic thesaurus compilation and automatic usage,term service, knowledge organization, and ontology construction are reviewed and prospected.According to the research findings of the existing literature, the domestic thesaurus research mainly focuses on the compilation, application, revision, and knowledge management of the comprehensive and professional thesaurus and the research on the terminology database. With the rapid development of information technology, the automation and networking of the compilation of the thesaurus, the diversification of the thesaurus application and knowledge services have become new trends.

2. Research Design

2.1. Data Procurement

The research uses China national knowledge infrastructure database(CNKI) advanced search tools, the search type is TI="thesaurus or subject thesaurus or index dictionary", and the periodical is limited to the information technology category of library information and digital and library, and the limited search time is 2010-2019 .The retrieval time is September 10th. A total of 530 related documents were retrieved, and 74 non-academic papers and irrelevant documents such as conference notices and solicitation notices were eliminated through manual sorting, and there were 456 valid journals remaining. The key fields of 456 documents were selected and analyzed.

2.2. Research Methods

The co-occurrence analysis method was first proposed by the French bibliometrics Callon in the middle and late 1970s, and its ideas are derived from the citation coupling and co-citation concepts of bibliometrics[6].Its principle is mainly to count the number of times they appear in the same document for a group of words, and perform cluster analysis on these words based on this, so as to reflect the close relationship between these words, and then analyze the co-occurring words the relationship between the themes represented to reveal the research hotspots and development trends of a certain research field[7].

After years of development, co-occurrence analysis has been widely used in many fields such as artificial intelligence, scientometrics, information science and information systems, information retrieval, and has achieved important research results.

3. Result Analysis

3.1. Trends in the Number of Papers

The distribution of research results in a research field over time can reflect the current development and research status of the field to a certain extent. Combined with figure 1, the papers downloaded from CNKI are distributed by age, and the thesaurus research papers are drawn out by year. The overall number of research papers fluctuates and declines, indicating that the current domestic attention to thesaurus research is not high, and the research progress is slow.

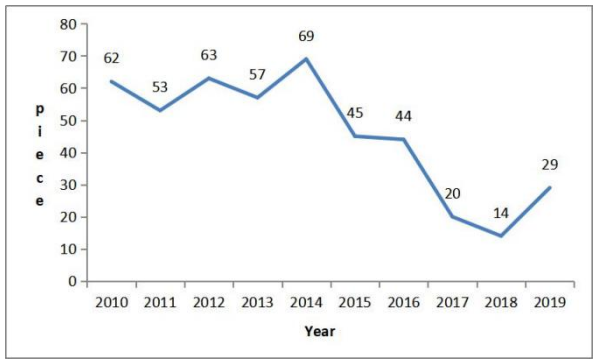

Figure 1: Number of paper published.

3.2. High-frequency Keyword Co-occurrence Network Analysis

The keywords in the literature are the concentration and refinement of the core content of the literature. The word frequency analysis method uses the frequency of keywords or subject words that can reveal or express the core content of the literature in a certain research field to determine the research hotspots and bibliometric methods of development trends. High-frequency keywords reflect the hot issues in this research field to a large extent. Statistical analysis of high-frequency keywords is intended to clarify the relationship between the focus, research topics and trends of a current research field. This study uses bibexcel to extract keywords and eliminate some meaningless keywords such as "meaning" and "standard" and merge keywords with the same meaning, such as "knowledge organization system" and "knowledge organization system", etc., complete the key word of cleaning work. In this study, 45 keywords with a frequency threshold ≥ 5 were selected for statistics(table 1:list the top 15)

Table 1: High frequency keyword table.

Serial Number	Keyword	Term Frequency
1	thesaurus	135
2	ontology	53
3	knowledge organization system	46
4	knowledge organization	40
5	classified chinese thesaurus	29
6	simple knowledge organization system	20
7	interoperation	18
8	correlative data series	17
9	sorting	16
10	subject indexing	16
11	organization of infromation	15
12	chinese thesaurus	14
13	semantice network	14
14	construction of thesaurus	12
15	term interrelations	12

These high-frequency keywords reflect the current research hotspots and research trends of the thesaurus to a certain extent. The research perspective of the researcher ranges from the thesaurus, subject indexing and retrieval method, sorting, ontology to knowledge discovery and digital library. It shows that domestic scholars in the field of library and information research have already used foreign research results to compile my country's first comprehensive vocabulary "Chinese Subject Thesaurus"; The development of subject indexing and retrieval method and sorting,which their controversy promote the process of integration of classification with thesaurus. With the development of network technology, the emergence and practice of concepts such as the semantic web have promoted the new direction of thematic thesaurus. The core keywords have high word frequency and relatively close relationship between words, which reflects the maturity of related research. The relationship between structural marginal words will change over time and may gradually become core keywords.

Although the high-frequency keywords can reflect the core and hot issues in the thesaurus research field to a certain extent, only the frequency of occurrence cannot fully reflect their internal relationship. The co-occurrence analysis method can be used to analyze the vocabulary in the literature or the co-occurrence of noun phrases is analyzed to further determine the relationship between the subjects in the subject represented by the collection. Using the co-occurrence analysis method, and with the help of Ucinet+Netdraw software, a map of important keywords in a certain research field can be constructed. In the graph, keywords are used as nodes, and lines are used to connect two keywords that appear in an article at the same time, and each link is given a different weight according to the importance of the relationship to distinguish the criticality of the keywords.Among them, the larger the keyword icon, the more central and the greater the influence of the keyword in the relevant research field. The thicker the connection between keywords, the more times the pair of keywords appear in an article at the same time, and the close connection between keywords. Use the Excel tool to further process the keywords selected by Bibexcel to obtain a 45×45 matrix, import the matrix into Ucinet, and obtain the keyword co-occurrence view through Netdraw (as shown in Figure 2)

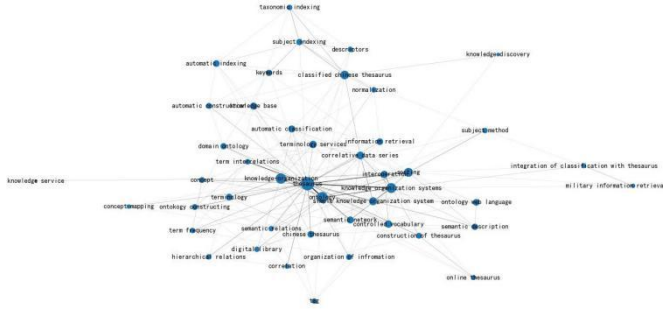

Figure 2: Knowledge map of co-occurrence.

It can be seen from Figure 2 and table 1: (1)The thesaurus is at the center of the research, with the highest centrality, and almost has a certain relevance to other keywords. This is because the thesaurus research mainly revolves around the thesaurus. (2) In addition to the thesaurus, the centrality of knowledge organization, ontology, and simple knowledge organization system is also relatively high, indicating that the current academic thesaurus is mainly based on these keywords, and the research system is relatively mature. Scholars can continue to dig on the basis of the predecessors and explore in depth. (3)Although keywords such as knowledge service and knowledge discovery appear on the edge of the co-word analysis graph, the degree of mutual relevance is not very high, but with the in-depth research on the thesaurus in academia, these problems are likely to be in the future and become a research hotspot in the middle of the year.

3.3. Clustering and Multi-dimensional Scale Analysis

SPSS is a statistical analysis software that integrates data management, statistical analysis, chart analysis, and output management. It is widely used in various disciplines. In this paper, system clustering and multi-dimensional scale analysis are used for research. System clustering method is a statistical method to find a statistical quantity that can measure the degree of similarity between these data or indicators based on a batch of data or indicators, and to draw a systematic classification map based on the closeness of all kinds of connections. Import the previously obtained matrix into SPSS, perform a systematic clustering analysis on it, and obtain a dendrogram. Compared with the cluster dendrogram,multi-dimensional scale analysis can intuitively infer that a certain research field is within the discipline in alower-dimensional space. With the help of dendrogram clustering results combined with multi-dimensional scale analysis, the vocabulary research is clustered into 3 groups, as shown in Figure 3.

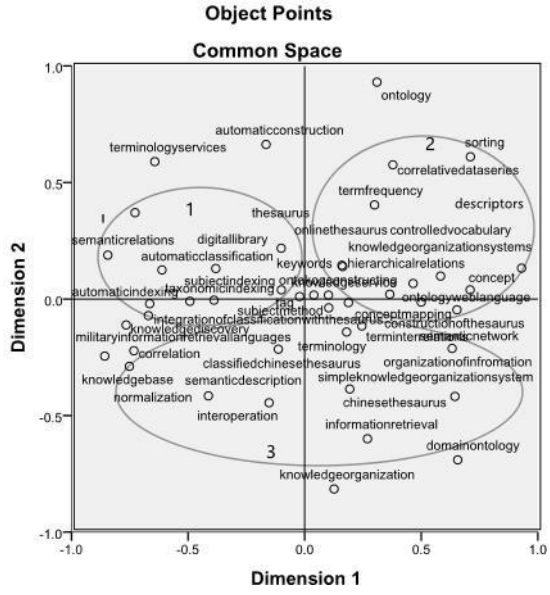

Figure 3: Clustering and multi-dimensional scale analysis.

It can be seen from Figure 3: (1)The first group is mainly devoted to information indexing[8], information indexing is divided into subject indexing and taxonomic indexing. Subject indexing is the process of assigning document subject identification based on specific topics. Subject indexing can use title language, descriptive language and keyword language, etc. Taxonomic indexing which is the process of assigning classification marks to documents based on specific classification language. In the big data environment, machines often need to extract keywords or classification numbers from the text that can express the content of the document information based on the relevant knowledge base, and use them for text retrieval and classification navigation.Therefore, the construction of knowledge base is one of the important research contents of automatic indexing.

(2)The members of the second group mainly focus on the research of information retrieval language, including the sorting, the subjectindexing and retrieval method ,and the integration of the classification with thesaurus to meet the development needs of the times. The web version of the "Chinese Classification Thesaurus", a major achievement of the integrated research on classification themes, was officially released in march 2010.

(3)The third group members mainly include the thesaurus, ontology and knowledge organization involve almost all fields of thesaurus research, which have a certain comprehensiveness and integrity. The rapid development of computer technology has promoted the computerization, management and maintenance of the thesaurus. Ontology theory is a hot issue in the field of library and information research in recent years. Ontology can reveal more complex semantic relations and describe concepts in formal language. Knowledge organization is the core issue in the field of library and information research. With the development of the semantic web, ontology theory and linked data, the use of SKOS language to semantically describe the thesaurus has become the mainstream, and the thesaurus will have new vitality in knowledge organization.

(4)The distribution of keywords in the same cluster is basically consistent with the results of systematic clustering. Moreover, the connection between cluster group 3 is relatively close, and it is relatively close to the center point, which shows that domestic scholars pay more attention to these research fields, and they often combine the two for comprehensive research. From this figure, we can also see that in the current thesaurus research, there are many fields involved, and the results are rich, with a certain dimensionand breadth. But the development of information technology continues to prompt researchers in the field of library and information to find new

constructions methods and models to meet the needs of social development.

4. Conclusion

According to the above series of analysis, it can be seen that the current thesaurus research is showing a trend of volatility and decline. Today, as information services are increasingly developed, users' demands for scientific and technological information are becoming increasingly diversified, which forces the builders of scientific and technological information service systems to constantly look for new ones construction methods and models, the thesaurus is increasingly hidden and backstage. However, no matter how it evolves, the thesaurus has not been abandoned as an effective knowledge standardization tool.It still plays an important role in the in-depth analysis of the subject area, synonym aggregation, and conceptual relationship recognition. Its development shows a diversified and diverse data format.The development trend of class table integration and the ontological transformation from terminology model to guiding concept model, and its basic functions and forms are also undergoing profound changes[9].

According to the results of the co-occurrence network analysis, we can see that the thesaurus research is mainly based on the comprehensive research of information organization. The development of the thesaurus shows networking, compilation automation and ontology transformation, and the release of professional vocabularies features such as linked data. Various thesaurus-based applications have rapidly increased, focusing on the automatic enrichment of vocabulary, the interoperability between different vocabularies, and the formation of domain knowledge services through thesaurus.

According to the results of co-occurrence cluster analysis and multi-dimensional scale analysis, we can divide the thesaurus research into three areas: comprehensive research on information organization, sorting and information indexing.This article is a summary of the 2010-2019 documents under the Library of Information and Digital Library, and explores the research status and direction of thesaurus in recent years. As some documents on other topics are not counted, it will interfere with the results.

References

[1] Aitchison J, Clarke S D.(2004) The thesaurus a historical viewpoint, with alook to the future. Cataloging Classification Quarterly,37,5-21.

[2] Ma Zhanghua, Hou Hanqing.(1999) Introduction to the Subject Method of Document Classification. Beijing: Beijing Library Press.

[3] Yu fengmin.(2014)A preliminary study on the research context of domestic subject list . Informationscience,32,12-17.

[4] Zhang Zhongqiu. (2012)A review of the research and practice of Chinese classification table and thesaurus in recent years. Library Work and Research,11,102-107.

[5] Zeng Jicheng, Yue Quan, Xu Huimin, Xu Wanying.(2017)The research course, hotspot and frontier of Thesaurus in China. Journal of agricultural books andinformation,29,83-85.

[6] Chu Jiewang, Guo Chunxia.(2011) The basic principle of co-word analysis and its implementation in EXCEL. Information Science, 29,931-934.

[7] Li Gang, Li Yuyao, Xie Zilin, Ba Zhichao.(2017) Research on the Influence of Mixed Keyword Selection Strategy on Co-word Analysis Effect. Information Theory and Practice, 40,110-116.

[8] Chen Baixue, Song Peiyan.(2018)TF-IDF-assisted indexing algorithm and empirical research based on user's natural annotation. Library and Information Service, 62,132-139.

[9] Liang Bing, Bai Haiyan, Wang Li, Qiao Xiaodong.(2016) The construction and application of the science and technology vocabulary of the National Science and Technology Book Documentation Center . China Science and Technology Resources Guide,48, 1-6+12.