SUBJECT ANALYSIS OF KNOWLEDGE EXTRACTION IN CHINA BASED ON K-MEANS ALGORITHM

Zhu Fu, College of Economics and Management, Nanjing University of Science & Technology, No.200 Xiaolingwei, Nanjing, China, 210094, 0086-25-84315963, fuzhu886@163.com
Yuefen Wang, College of Economics and Management, Nanjing University of Science & Technology, No.200 Xiaolingwei, Nanjing, China, 210094, 0086-25-84315963, yuefen163@163.com
Peng Guan, College of Economics and Management, Nanjing University of Science & Technology, No.200 Xiaolingwei, Nanjing, China, 210094, 0086-25-84315963, guanpeng1983@163.com

ABSTRACT

The aim of this study is to map the intellectual structure of knowledge extraction (KE) field in China during the period of 2004-2013. Co-word analysis is employed to reveal the patterns of KE field in China through measuring the association strength of keywords in relevant journals. Data is collected from Chinese Journal Full-Text Database during the period of 2004-2013. And then, 103 keywords with proper frequency are selected as concepts from related articles of KE in China, which are classified into five clusters based on K-means algorithm after keyword standardization. Finally, the main contents of these five subjects are discussed, respectively.

INTRODUCTION

Recently, knowledge extraction (KE) is becoming a new research focus. It mainly refers to extracting knowledge units from the text. There are many similar or related concepts about KE, including data mining, knowledge discovery, knowledge acquisition, information extraction, information acquisition, etc. These concepts may seem similar, but in fact, they are different. Definition of the concept is the premise and basis for scientific research, how to conduct classification and segmentation for these concepts is one of the important works for conducting research [1]. Based on this point, we use the classic K-means algorithm to classify the relevant concepts on KE, and then analyze its main content according to the classified topics.

METHODOLOGY

This study assumes that the keywords with proper frequency are chosen as the subject to represent the specific topics. It indicates that any two keywords, co-occurring within an article, are relevant in the topics which they refer to [2]. The presence of many co-occurrences of pair of keywords within articles demonstrates that they may belong to one research theme [3]. In an article, keywords are words or phrases selected to reveal the subject and meet the retrieval need, which is the expression of core content, ideas and argumentation methods [4]. The keywords could provide adequate description of a paper's content. The readability and effectiveness of keywords in an article can be guaranteed because they are given by authors who have a professional background. Therefore, we will use keywords as the research object of this experiment.

Relation measure between words used in the field of library and information science, mainly including inclusion index, proximity index, the equivalent coefficient, Salton's Cosine coefficient, etc [5]. To eliminate the effects of the difference of absolute value in the original co-occurrence matrix, we use Salton's Cosine coefficient (S_{ij}) [6] to calculate the correlation between words, the coefficient is defined

as follows:

$$S_{ij} = C_{ij} / (C_i \times C_j)^{1/2} \quad (1)$$

Where C_i is the occurrence frequency of keyword i in the set of articles; C_j is the occurrence frequency of keyword j in the set of articles; C_{ij} is the number of documents in which the keyword pair appears (keyword i and keyword j). S_{ij} has a value between 0 and 1. S_{ij} measures the probability of word i appearing simultaneously in a document set indexed by word j, and inversely, the probability of word j if word i appears, given the respective collection frequencies of the two words.

Fast and high-quality terms clustering algorithms play an important role in our work [7], as a result, K-means algorithm is exploited to cluster the keywords selected.

DATA COLLECTION AND PROCESS

In our work, once a research area is selected, keywords are extracted from the related journal articles; and then, a correlation matrix will be built.

Data Collection

We choose Chinese journal full-text database (CJFTD) as data source and retrieve the related articles about KE field from CJFTD with a search strategy as follows: title = knowledge extraction or keyword = knowledge extraction and core journals, the time span from 2005 to 2014 (retrieval date: June 5, 2015). Notice that, in this retrieval, the category of data source is "core journals". These journals, embodied in Chinese science citation database (CSCD) or Chinese humanities and social science citation database, are important and leading in a special research area in China.

Moreover, in order to exclude the interference of irrelevant documents to ensure the precision and recall, articles without keywords, notices of meetings, book reviews, editorials, meeting abstracts, newsletters or notes are excluded. Finally, 4,169 articles are selected as the co-word analysis sample.

Data Process

In total, 21,218 keywords are collected from the selected 4,169 articles, and 10,467 unique keywords are extracted from all keywords. The average number of keywords per article is found to be 5.09. There are many forms of keyword expression in the keywords list, such as synonyms, abbreviation, mixed case, broad/narrow, etc. Thus, the various keywords and phrases are standardized by selecting an appropriate heading (see Table 1) and excluding the general terms which are meaningless or too broad (e.g., theories, construction, development, applications, information, knowledge). After keyword standardization, the top 20 high-frequency keywords are listed in Table 2.

Table 1: Some examples of keyword standardization

Heading	Original keyword
ontology	ontological, domain ontology, knowledge ontology, role ontology, product ontology, event ontology, resource ontology, bridge ontology
database	network database, spatial database, process database, comprehensive database, ancient databases, relational databases, commercial literature databases, thematic databases, public databases, biological databases, full-text databases
KDD	knowledge discovery in databases, knowledge acquisition in database, knowledge discovery based on database; knowledge base discovery
knowledge	knowledge expression, knowledge representation of historical figures, interpreted knowledge

representation	representation
rough set	ROUGH set, ROUGH SET, rough set theory, fuzzy rough sets, variable precision rough sets, multiple rough set, the probability of rough set, multiple probability rough set, gray rough set, dominance relation rough set, multi-precision rough set
data mining	spatial data mining, distributed data mining, web data mining, extension data mining, medical data mining, text mining, incremental data mining, web data mining, multidimensional data mining, temporal data mining

Table 2: The top 20 high-frequency keywords

No	Keyword	frequency	No	Keyword	frequency
1	Knowledge Acquisition	865	11	Knowledge Management	126
2	Knowledge Discovery	787	12	Expert System	120
3	Information acquisition	760	13	Library	120
4	Information Extraction	599	14	Enterprises	93
5	Data Mining	381	15	Network	92
6	Knowledge Mining	207	16	Student	90
7	Rough Set	190	17	Database	80
8	Rules	174	18	Knowledge Representation	79
9	Ontology	155	19	Fault Diagnosis	77
10	Algorithms	149	20	Knowledge Base	74

EXPERIMENTS

Concept Clustering

The 103 keywords with the frequency more than 10 are chosen. Then, the responding 106× 106 correlation matrix is constructed (see Table 3). These keywords can be clustered into five clusters which include different number of keywords. K-means clustering analysis using the "Iterate and classify" is conducted on the correlation matrix to complete cluster sample analysis by using the software SPSS19.0. Each cluster is to be given a name that can describe its unique characteristics. We automatically select three keywords which have the highest frequency value to represent the cluster name. Clustering results are shown in Table 4.

Table 3: Correlation matrix (fragment)

Keyword	Knowledge Acquisition	Knowledge Discovery	Information acquisition	Information Extraction
Knowledge Acquisition	1.0	0.006060	0.0	0.008335
Knowledge Discovery	0.006060	1.0	0.002586	0.005825
Information acquisition	0.0	0.002586	1.0	0.008892
Information Extraction	0.008335	0.005825	0.008892	1.0

Table 4: Five clusters of KE in China

Cluster	Cluster name	Keywords
1	knowledge acquisition, expert systems, knowledge representation	Knowledge acquisition, expert systems, knowledge representation, fault diagnosis, knowledge base, neural networks
2	Knowledge discovery, data mining, models	Knowledge discovery, data mining, models, algorithms, applications, databases, clustering, information processing, KDD, artificial intelligence, natural language processing, data warehousing, visualization, customer relationship management, non-related literature, data analysis, data mining technology, computer
3	Information extraction, ontology, knowledge extraction	Information extraction, ontology, classification, information discovery, search engines, knowledge extraction, XML, text mining, WEB, wrapper, DOM, templates, support vector machines, semantic, WEB mining, CRF, the Semantic Web, metadata, information Get channels, e-commerce, named entity recognition, pattern matching,

		review, cluster analysis, semantic annotation, HTML, information management, information organization, knowledge discovery systems, information fusion, microblog, Internet, personalized service, regular expressions, decision Support Systems
4	knowledge management, enterprise, knowledge services	Knowledge, knowledge mining, knowledge management, enterprise, networking, knowledge sharing, decision making, knowledge services, knowledge engineering, social capital, knowledge organization, strategic alliances, knowledge transfer, information technology, knowledge creation, competitive advantage, knowledge, reasoning, social networks knowledge element, the knowledge economy
5	Information acquisition, Internet, information services	Information acquisition, Internet, the way to acquire information, IT, information technology, information resources, information services, the channels of information acquisition, users, information needs, the digital, networked environment, traditional media, access to information technology, factors, trends, information literacy , information behaviour, personalization, information age, new media

Cluster Analysis

Current research structure of KE in China is analyzed and explained according to the clustering results, as follows:

(1) Knowledge Acquisition (KA). KA is one of the three steps (including knowledge acquisition, knowledge representation and knowledge utilization) of knowledge engineering, and KE is the most effective way and the critical path of knowledge engineering because KE is an effective way of KA. Moreover, KA is a prerequisite for knowledge representation, in which neural networks is the classic method. Knowledge is stored in knowledge base after KA. Knowledge base is the source of expert systems and can be used in the field of product design, fault diagnosis applications, etc.

(2) Knowledge Discovery (KD). KD is the same as data mining, which involves database technology, artificial intelligence, cognitive science, etc. KD is used to help business decisions in the field of knowledge management, and it focuses on the implementation of data mining technology in information systems and the promotion of practical activities in business operations, especially in customer relationship management and e-commerce. KD in the field of NLP is used to discover implicit knowledge in the information carrier(e.g., non-relevant document, database, data warehouse) with methods and techniques for example, models, algorithms, visualization, data analysis. The nature of KD is an artificial intelligence (AI) technology mainly used in intelligence science and AI theory is its theoretical basis.

(3) Information Extraction (IE) and knowledge extraction (KE). IE is the basis of information organization and information management. IE and KE are very similar and they can be distinguished by information and knowledge. It is difficult to make a distinction between IE and KE completely because the methods and techniques of KE are mostly derived from IE. Information discovery is a part of IE process, network is the main target of IE, and web mining is a form of network information extraction, in which microblog is one of the major carriers of network information dissemination and gathering in recent years. XML and metadata is the basis element of information organization, information semantization is one of knowledge organization technologies. IE technology is abundant, including ontology, DOM, wrapper, support vector machine, conditional random filed, pattern matching, regular expressions, templates, cluster analysis, etc. IE is commonly used to support search engine optimization, named entity recognition, personalized service, semantic annotation, decision support systems and other applications.

(4) Enterprise Knowledge Management (EKM). Innovation management is the core of knowledge

management (KM). The basic goals and tasks of KM are to improve the resilience and innovation capacity of organizations. In the time of knowledge economy, the main content of EKM is to conduct knowledge mining, knowledge sharing, knowledge organization and knowledge transfer within the enterprise, and to provide knowledge services for other corporate or individual outside of it, with the ultimate aim is to create and reuse knowledge. Innovation is not only the source of economic growth, but also the critical for core business competitiveness. Information construction is the basis condition of KM, and networking is one of the supporting techniques for information constructions. Knowledge-based enterprises are knowledge-centric in business activities. EKM is to develop knowledge resources of enterprise as a starting point and to obtain a competitive advantage as a final goal. Management scholars generally believe that company should be managed with the theory and methods of KM in the process of innovation.

(5) Information Acquisition (IA). Today is the information era of Internet-driven, the way and channel of IA both have larger changes compared to the previous. It is difficult to meet the information needs of current users because information dissemination in the past is mainly based on traditional media. In the environment of Internet, information mainly obtained from network in which there is a wealth of information resources. Internet is becoming the new media of information dissemination, which provides users information services with information technology. Because of Internet, the user's information literacy also will be changed and a growing trend of personalized information behaviour will be presented.

CONCLUSION

In this paper, keywords are extracted from the relevant paper as the concepts about KE, and then Slaton's Cosine coefficient is used to calculate the relations between keywords. After the correlation matrix is built, these keywords are clustered into five clusters by using K-means algorithm for depth study of research structure of this topic. Finally, five subjects are analyzed to provide a theoretical basis for the further work of KE.

ACKNOWLEDGMENTS

This research is supported in part by National Natural Science Foundation of China (grant No. 71373124).

REFERENCES

[1] Hua, B. L. & Zhang, X. M., Describing character and trend of knowledge extraction based on discussion on concepts relative with knowledge extraction. *Information Science*, 2010, 28(2), 311-315.
[2] Cambrosio, A., Limoges, C., Courtial, J. P. & Laville, F., Historical scientometrics? Mapping over 70 years of biological safety research with co-word analysis. *Scientometrics*, 1993, 27(2), 119-143.
[3] Ding, Y., Chowdhury, G. & Foo, S., Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing and Management*, 2001, 37(6), 817-842,.
[4] He, Q., Knowledge Discovery through co-word analysis. *Library trends*, 1999, 48(1), 133-59.
[5] Feng, L. & Leng, F. H., Development of theoretical studies of co-word Analysis. *Journal of China Library*, 2006, 32(2), 88-92.
[6] Tan, P. N., Steinbach, M. & Kumar, V., *Introduction to data mining*. Boston: Pearson Addison Wesley, 2006.
[7] An, X. Y. & Wu, Q. Q., Co-word analysis of the trends in stem cells field based on subject heading weighting. *Scientometrics*, 2011, 88(1), 133-144.