Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Environmental Sciences 10 (2011) 561 – 567

Procedia
Environmental Sciences

2011 3rd International Conference on Environmental
Science and Information Application Technology (ESIAT 2011)

Research on Field Characteristics of Shared Earth System Science Data Using Keyword Analysis and Visualization

Runda Liu[a], Haiqing Lin[b], Hui Zhao[a] a*

[a] *Institute of Scientific & Technical Information of China, Beijing, China*
[b] *School of Management and Economics, Beijing Institute of Technology, Beijing, China*

Abstract

This paper conducts analysis on the metadata information of shared Earth System Science data of China Scientific Data Sharing Project (SDSP) published online by Portal of Chinese Science and Technology Resource. High-frequency keywords are extracted from the metadata, keyword clusters are formed by means of co-occurrence matrix and K-core analysis, and then network visualization is made by Pajek software. This research reveals the field characteristics of the data through analyzing the keywords network structure, make conclusion on the hot acquisition spots of data in Earth System Science field shared on SDSP.

*Keywords:*Scientific Data; Earth System Science; Co-occurrence Analysis; Social Network Analysis;

1. Introduction

Scientific Data are data resources relating scientific research, specifically refers to basic sci-tech data, materials generated by social sci-tech activities and data products and related information generated for different requirements [1]. Scientific data sharing means sharing scientific data among scientific research groups and individuals by certain strategies and techniques, it has become an important guarantee for the smooth evolution of scientific research, and it effectively promotes the sustainable development and innovation of contemporary science and technology.

* Corresponding author. Tel.: +86-10-58882532; fax: +86-10-58882532.
E-mail address: liurd@istic.ac.cn

doi:10.1016/j.proenv.2011.09.091

Worldwide, series of studies and practices on scientific data sharing methods based on computer network were conducted in developed countries and related international organizations [2]. In China, Scientific Data Sharing Project (SDSP) under National Science & Technology Infrastructure program (NSTI) is the main project of scientific data sharing activities implemented in sci-tech community [3]. In September 2009, Portal of Chinese Science and Technology Resource which is a portal of NSTI was launched and started providing service to the whole society [4], it gathers superior resources and achievements of NSTI so that general public can download various kinds of scientific data from it. The data cover the fields of medical hygiene, advanced manufacturing, earth system, geology & minerals, forestry, meteorology, communication, marine, agriculture and earthquake.

At present, metadata information of scientific data under the Portal of Chinese Science and Technology Resource is open for share. Data characteristics can be rediscovered by the analysis of the metadata information. However, there are few analysis cases on these data resource characteristics based on data attributes, especially based on keywords in the metadata. This article adopts keyword co-occurrence network method and uses visualization technology to analyze the field characteristics and hot acquisition spots of the earth system science field data in SDSP, and then propose analysis conclusions and suggestions to provide references for researchers and scientific data sharing activities.

2. Data sources and analysis methods

2.1. Data sources

International organization for standardization (ISO) regards metadata as the description of data content, quality, conditions and other characteristics [5]. The data used in this research are metadata of the scientific data from Earth System Science field data resources of China Scientific Data Sharing Project published online by Portal of Chinese Science and Technology Resource. Data acquisition time was in October 2010. 736 metadata were obtained. These metadata are originally created by Earth System Science Data Sharing Network [6] which is a project under SDSP. Earth System Science Data Sharing Network started collecting and integrating data resource in 2003, the amount of data accumulation rise consequently in the following years.

A new keyword field used for analysis was set by extracting and perfecting terms from the original keywords and abstract fields, the process includes the merger of synonyms or similar words and the removal of common words and noise words etc. Then new keywords field for analysis were revised, audited and verified by field experts to ensure data validity. The resulting new keywords field for each metadata item contains 2~5 new keywords separated by comma.

After statistics, there are 1046 different keywords in the new keywords field. High-frequency keywords are selected according to their cumulative frequencies. To make a focus for this study, we select top 51 keywords with at least 10 times of appearance shown in table 1.

Table 1. High-frequency keywords of Earth System Science field data

No.	Keywords	Frequency	No.	Keywords	Frequency	No.	Keywords	Frequency
1	China	160	18	Temperature	17	35	mud-rock flow	12
2	south pole	146	19	Russia	16	36	ACT	11
3	Zhongshan station	66	20	black soil	16	37	earth tide	11
4	Qinghai-Tibet plateau	64	21	Chengdu	15	38	Hailun	11

5	distribution	47	22	loess plateau	15	39	United states	11
6	soil	43	23	digitalize	15	40	air temperature	11
7	northeast	32	24	Yunnan	15	41	hydrology	11
8	sky survey	27	25	minya konka	14	42	Tibet	11
9	north pole	26	26	land utilization	14	43	observation	10
10	Fujian	25	27	location	14	44	seawater	10
11	Prydz bay	22	28	Changcheng station	14	45	forest	10
12	ground Meteorology	21	29	Chongqing	14	46	social economy	10
13	remote sensing	20	30	resource	14	47	projection	10
14	routine meteorological observation	17	31	DomeA	13	48	west	10
15	spatial distribution	17	32	climate	13	49	salinity	10
16	meteorology	17	33	statistics	13	50	Yanting station	10
17	Sichuan	17	34	Guizhou	12	51	gravimeter	10

2.2. Analysis methods

These high frequency keywords reflect the hot acquisition spots of the Earth System Science field data. In order to further reveals the relationship between these high frequency keywords, co-occurrence and social network analysis is adopted.

In the 1970s, French bibliometrics experts first create the concept of co-occurrence analysis. Co-occurrence analysis, as the basic method of content analysis, is a kind of data mining technology with combination of quantitative and qualitative analysis, it mainly applied to scientific literature field by calculating the frequency of the occurrence of two words in the same article by statistics, and then clustering these words to reveal their affinity-disaffinity relationships so as to analyze the structural changes of the disciplines and subjects which they represent [7], this can well reflect current academic research hot spots, knowledge structure and development trend in certain disciplines. One major ways of co-occurrence analysis is to determine the concept map or knowledge network structure among these representative terms and can describe the subjects of a discipline in details through a series of similar maps [8]. We can analyze data structure in three aspects by co-occurrence analysis: (1) network of vertexes and edges, the vertices represent those representative terms and one edge represents the relationship between two terms, which form a network; (2) the distribution situation of the network in the interaction of the vertexes; (3) the dynamic change of the network in time sequence. We apply keywords co-occurrence analysis method to scientific data field to form network and analyze its distribution.

What's more, In order to show the relationships of these keywords more clearly, in the next part of this work, we first introduce a social network analysis method called K-core analysis, a concept first proposed by Seidman. K-core network is a particular network of a given network, where every vertex at least has connection with other k vertexes in a given core. With K-core analysis, we can get different networks by changing the value of K. Along with the increase of K, the number of members in the K-core set will gradually decrease, and the relationships of the members will be closer, so the network characteristics can be seen even apparent.

3. Field characteristics analysis

3.1. Field characteristics based on co-occurrence analysis

It needs to find out the concurrence times of two keywords to construct the co-occurrence matrix before clustering calculation in traditional co-occurrence analysis. The calculation of the co-occurrence times of keyword T_j and $T_{j'}$ is defined as:

$$\omega_{jj'}=\begin{cases}\sum\limits_{i=1}^{n} \varepsilon_{ij}\varepsilon_{ij'} & if \quad j \neq j' \\ 0 & if \quad j \neq j' \end{cases} \quad 1 \leq j \leq n\,,\ 1 \leq j' \leq n\,.$$

$\omega_{jj'}$ represents the concordance relationship between keyword T_j and $T_{j'}$.

And $$\varepsilon_{ij}=\begin{cases}1 & if\ the\ jth\ keword\ is\ in\ the\ ith\ article \\ 0 & if\ not\end{cases}$$

It just reflects superficial phenomenon by above co-occurrence relationship, because the concurrence frequency of two keywords is directly affected by the respective frequency of each one [9], which does not reflect their real affinity-disaffinity relationship. To reflect the affinity-disaffinity relationship between keywords, a scholar has proposed mutual inclusive coefficient method whose statistical formula is: $E_{ij}= C_{ij} \quad C_i \qquad C_{ij} \quad C_j = C_{ij}{}^{2} \qquad C_i*C_j$, where C_{ij} represents the concurrent frequency of the keyword and , C_i represents the frequency of the keyword , C_j represents the frequency of the keyword in the literature set, and E_{ij} represents the mutual tolerance values of the keyword and , which is between 0 and 1. Mutual tolerance value represents the interdependence degree between two keywords. Theoretically, the closer the interdependence between two keywords the bigger chance for them to be clustered together [10].

For the above 51 high-frequency keywords, we pairwise compile statistics on their concurrent times in the same metadata. In co-occurrence analysis, the higher the co-occurrence frequency of two keywords, the closer their relationship. Use excel macro function to count concurrence times of the 51 keywords which can be regarded as matrix elements to form a 51x51 matrix. Use mutual tolerance coefficient method to process this matrix to obtain the mutual tolerance coefficient matrix.

In the vast node network, if we don't filter connection weights of the network lines, there will be full of lines, and we will be interfered to find out relative important (with greater weight) relationships in the network by some lines with relatively minor weight. Therefore, it is necessary to set up a threshold for the connection weights. The lines with weight less than the threshold will not be shown in the network. In the mutual tolerance coefficient matrix, we eliminate the elements with the value of 0 and then calculate the average value of the matrix elements, which is 0.078948 here, as the average mutual tolerance coefficient value. According to this threshold we process the mutual tolerance coefficient matrix with value greater than or equal to the threshold are set to 1, otherwise 0. In the result, we get a new matrix with element value of 0 or 1. The affinity-disaffinity relationship of the keywords reflected by the new matrix is very close. Based on this, we try to show the closest affinity-disaffinity relationship of these keywords. The new matrix is saved in file format which can be import into Pajek [11] for visualization.

The co-occurrence network of the high-frequency keywords of the Earth System Science field data is drawn by using Pajek according to the binary matrix as shown in figure 1. There are 9 clusters which include at least 2 nodes. The size of the node shows its degree. The greater degree the bigger the size of the node, which means there are more other nodes connected to it, and it is supposed to be the core node of the cluster.

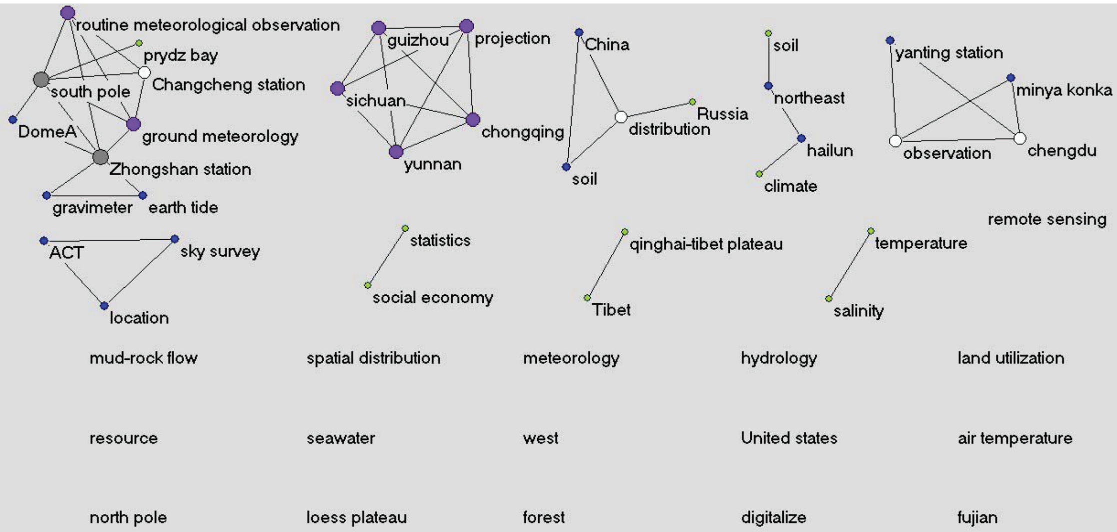

Fig. 1. Co-occurrence network of the high-frequency keywords

The clustering and visualization result shows the distribution of hot acquisition spots of current Earth System Science field data in SDSP. From the research field angle, top data acquired is data about polar research, then meteorological observation, projection, soil and distribution, climate, sky survey, social economic statistics subsequently. Taking the polar research for example, keywords in this cluster are Changcheng station, ground meteorology, routine meteorological observation, South Pole, Zhongshan station, DomeA, and gravimeter tidal. These keywords reflect the hot acquisition spots of the polar research data. The keywords such as South Pole, Zhongshan station, ground meteorology and routine meteorological observation are the core keywords according to their node size, this shows that Zhongshan station in the South Pole is the key data acquisition area in polar research and meteorological observation data are the main data collected.

From the regional angle, the shared data is mainly collected in the Antarctic, western and northeast region of China, Tibet and Qinghai-Tibet plateau region. The data of the Antarctic are mainly from the Antarctic Zhongshan, Changcheng station and Prydz Bay. The western region is the hot area of the Earth System Science field data acquisition. There are two clusters about western region and Sichuan province is hot data acquisition area in the western region. As shown in the figure, there are 16 isolated nodes. These keywords are: remote sensing, debris flow, spatial distribution, meteorology, hydrology, land use, resources, water, western, United States, temperature, arctic, loess plateau, forest, digital and Fujian province. It is the result of the average tolerance coefficient threshold value setting.

3.2. Network analysis based on K-core

The max value of K is 4 in the co-occurrence network of selected keywords, which means that when the value of K is 4, all nodes in the K-core at least connect with the other 4 nodes in the same core, and namely the mutual tolerance coefficient values of these keywords are equal to or greater than the average mutual tolerance coefficient value. Therefore, we can regard 4-core as the sub-network with the closest relationships of the high-frequency keywords co-occurrence network of the Earth System Science field data. The 4-core network is shown as figure 2 (a). As shown in the figure, only one cluster exits, keywords represented by nodes are Sichuan, Yunnan, Chongqing, Guizhou, and projection, the first four keywords are province name, *projection* is field keyword and others is region keywords. It might because that projection data or data projection is hot spot in current Earth System Science field data acquisition in Sichuan, Yunnan, Chongqing and Guizhou.

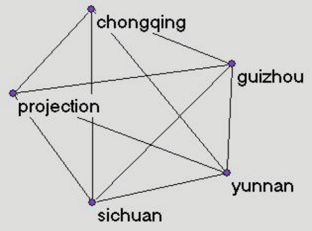

Fig. 2 (a) 4-core network;

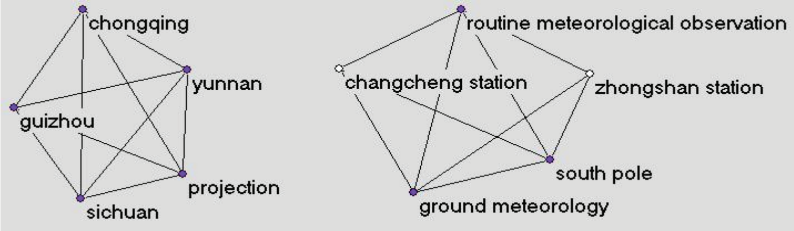

Fig. 2 (b) 3-core network

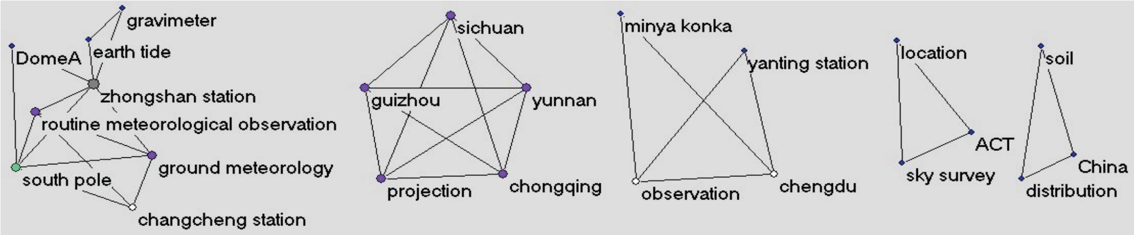

Fig.2 (c) 2-core network

In order to present more in-depth research on data acquisition hot spots, we introduce keywords with the secondary closer relationships by gradually reducing K value. 3-core is shown as figure 2 (b) which adds new nodes based on figure 2 (a). It is interesting that these new added keywords nodes form a new cluster without any connection with the original cluster. These new added keywords show that the hot acquisition spot of the Antarctic data is meteorological observation data, which further verifies previous view. In 2-core network shown in figure 2 (c), the new added keywords nodes labeled by DomeA, earth tide and gravimeter are connected with the cluster including keywords meteorological observation and South Pole, which means that ice vault investigation, earth tide observation and gravimeter survey are secondary hot fields in the Antarctic data acquisition, Besides the rest of the new added keywords form independent clusters, this shows that the type of our Earth System Science field data shared are various at present.

4. Conclusion

This paper carries on deep analysis on the keywords in the metadata information of Earth System Science field data resources of China Scientific Data Sharing Project published online by Portal of Chinese Science and Technology Resource by using co-occurrence network and visualization. We find that at present the shared Earth System Science field data are mainly concerning about the fields of polar research, meteorological observation, projection, soil and distribution, climate, sky survey, social economic statistics, etc. Data collection areas mainly concentrate in the Antarctic, the western and northeast region of China, Tibet and the Qinghai-Tibet plateau region. The hotspots of data acquisition are the projection observation data in the western region and Antarctic meteorological observation data.

The sharing of scientific data in Earth System Science field is still in the primary stage. For the fact that Earth System Science field data involves multiple research contents and are from different regions, it is one of the important tasks for future scientific data sharing to effectively manage and share the data; also it is very significant to forecast the long-term trend of scientific data acquisition in the future. However, in this research, because of the limited quantity of the shared scientific data and big difference between data distribution in the time dimension, it is still unable to predict the long-term development trend, this is one direction for further research and using network and visualization analysis will be very helpful during the cause.

Acknowledgements

This work is supported by ISTIC project: S&T Information Resources Integration, Sharing and Comprehensive Service System Construction (ZD2010-7-5) and preliminary research fund of ISTIC (YY-2010013). We acknowledge the Scientific Data Sharing Project from where data of this work were acquired.

References

[1] Huang Dingcheng, Guo Zengyan. *Scientific data sharing management research*. Beijing: China science and technology press;2002 : 389.

[2] Liu Runda. Scientific Data Portal& its Building Practices--Earth System Science Data Sharing Network as an Example. Ph.D. dissertation of Chinese Academy of Sciences; 2009.

[3] http://www.scientificdata.cn: Scientific Data Sharing Project.

[4] Ministry of Science and Technology. The Open of Portal of Chinese Science and Technology Resource. URL: http://www.most.gov.cn/kjbgz/200909/t20090927_73398.htm; 2009.

[5] Zhang Yingjun, Xie Bingong, Guo Yongyi. The Application of Metadata Technology in Scientific Data Sharing Platform. Journal of Taiyuan University of Technology; 2009,vol. 40(4), p.341-344.

[6] http://www.geodata.cn/: Portal of Data Sharing Infrastructure of Earth System Science.

[7] Feng Lu, Leng Fuhai. Development of Theoretical Studies of Co-Word Analysis. Journal of Library Science in China; 2006, vol.32(2), p.88-92.

[8] Liu Zeyuan, Yin Lichun. Visualization of International Science of Science Co-word Network. Journal of the China Society for Science and Technology Information; 2006,vol.25(5), p.634-640.

[9] Kang Yuhang, Su Jingqin. Visualization of Technology Tracking Based on Co-Word Analysis: An Empirical Study of Highway Engineering. Journal of the China Society for Science and Technology Information; 2008,vol.27(4), p.566-571.

[10] Li Jia. Role of co-word matrix in analysis of clustered words. Chin J Med Libr Inf Sci; 2009,vol.(4), p.77-80.

[11] http://pajek.imfm.si: Pajek—Program for Analysis and Visualization of Large Networks.