Research on the Clustering Method of Agricultural Scientific Data Based on the Author's Scientific Research Relationship

Dingfeng Wu, Liyun Wang, Jian Wang, Hua Zhao, Guomin Zhou

Research on the Clustering Method of Agricultural Scientific Data Based on the Author's Scientific Research Relationship

Dingfeng Wu, Liyun Wang(✉), Jian Wang, Hua Zhao, Guomin Zhou

Agricultural information institute of CAAS, Beijing 100081, China

{wudingfeng,wangliyun,wangjian02,zhaohua,zhaoguomin}
@caas.cn

Abstract. Focusing on semantic parse and bias problems during the clustering process of agricultural scientific data, a clustering method for agricultural scientific data based on author's scientific research relationship is proposed in this paper. Meanwhile, an assessment algorithm of the scientific research relationship based on co-author ship and authors′ inter-citation is put forward. Finally, the experimental results proved that the proposed clustering method for the agricultural scientific data can effectively improve error classification caused by semantic parse and bias.

Keywords: Scientific data· Data clustering· Scientific research relationship.

1 Introduction

Cluster analysis for the scientific data is a class of important calculation in the process of storing, processing and displaying agricultural scientific data. Cluster analysis for scientific data provides the basis for the establishment of scientific data storage system, in favor of effective management of scientific data, but also it can be the basis for the improvement of browsing and retrieval process of users, in favor of a substantial increase in browsing and retrieval efficiency of users [1, 6].

At the present stage, few studies are on research of cluster analysis for scientific data. The methods mainly based on the fields of scientific data and the keywords of metadata. However, this method has two following problems: first, ambiguity and incongruous granularity is the widespread problem in the description of the subject area, and a lot of scientific data often across multiple disciplines; second, a certain subjective randomness exists in the generating process of keywords. Many keywords are ambiguous, and conceptual keywords also have granularity problem. The two above issues illustrate that there exists a gap between the semantic description and semantic

* Agricultural information institute of CAAS, Beijing, China

entities, as well as it faces extremely complicated semantic parse problem in the description of scientific data using the text or conceptual tags [2], [7], [8], [12.13.14]. Therefore, there is a semantic bias problem difficult to solve in the Cluster analysis method for the scientific data based on them, which, which affects the density and within-class relativity degree of the clustering results.

The author of scientific data is a metadata item that represents a unique clear entity. It is not ambiguity in the semantic representation. Authors of scientific data are mostly engaged in scientific researches, and have strong consistency and continuity in research subject or field. Authors of scientific data are mostly subject or field. The scientific data produced by them are highly correlated with the author's research subject and field naturally. If two authors have a close relationship [3], it usually means that there is a correlation between their research fields. So their scientific data are likely correlated. It is a hopeful to avoid the complex semantic problems and get high quality clustering results using clustering method based on author's scientific research relationship.

The agricultural science data sharing Center is determined by the Ministry of science and technology, "national science and technology basic conditions platform" supporting data center for the construction of one pilot. 373TB [4] of agricultural scientific data is stored in this center. In this study, we try to analyze authors' research relationship network based on the rich resources of the center, then related scientific data is clustered based on the closeness degree of the authors' research relationship.

2 Materials and methods

2.1 Analysis of the research relationship

Analysis of the academic relationship mainly from two aspects, academic cooperation relationship and the mutual quotation relationship [5], [9], [10], [11]. In this study, the two aspects are analyzed on authors of 1700 scientific data selected, in order to quantify the degree of academic connection between the two authors. Fig.1 shows partial data for the academic connection of these authors.

Author	Number of citation	Number of cited	Number of cooperation
Zuo Tao	7	7	18
Zhou Xiaorong	6	1	9
Zhu Lixin	2	10	4
Liu Hongchun	1	1	1
Liu Jing	13	5	4
Zhu Yonggui	1	10	2
Zhou Qiubai	1	3	1
Fu Zidong	1	2	1
Ding Tianming	14	1	1
Dan Zhixin	2	7	2

Fig.1 Partial data for the academic connection of these authors

In the analysis of academic cooperation relationship, two authors' relationship is rated on a scale of zero to ten by analyzing the historical cooperation data of the author's. This score represents the degree of importance in each other's scholarly circles in the form of academic cooperation. So this score of author *A* and author *B* which is denoted as S_H, can be split into the degree of importance of

author *A* in author *B*'s cooperation circles which is denoted as S_{AB} and the degree of importance of author *B* in author *A*'s cooperation circles which is denoted as S_{BA}. The algorithm is:

①Figure out that author *C* is the one author *A* worked with most of all among these 1700 authors, their cooperative time is denoted as C_{maxA}。

②Figure out that author *D* is the one author *B* worked with most of all among these 1700 authors, their cooperative time is denoted as C_{maxB}。

③Figure out the cooperative time of author *A* and author *B*, denoted as *C*。

④Calculate author *A* in author *B*'s cooperation circle S_{AB}

$$S_{AB} = \frac{C}{C_{maxB}} * 10 \quad (1)$$

⑤Calculate author *B* in author *A*'s cooperation circle S_{BA}

$$S_{BA} = \frac{C}{C_{maxA}} * 10 \quad (2)$$

⑥Calculate the score of author *A* and author B: S_H

$$S_H = \frac{1}{2}S_{AB} + \frac{1}{2}S_{BA} \quad (3)$$

When analyzing the mutual quotation relationship, it is divided into two categories: the citing relationship and the cited relationship. Two categories are rated on a scale of zero to ten, they are S_Y and S_Y' respectively. The score of the mutual quotation relationship $S_{HY} = \frac{1}{2}S_Y + \frac{1}{2}S_Y$'.

This score S_Y of author *A* and author *B* represents the degree of importance in each other's scholarly circles in the form of quotation. So S_Y can be split into the degree of importance of author *A* in author *B*'s citing circles which is denoted as S_{AB} and the degree of importance of author *B* in author *A*'s citing circles which is denoted as S_{BA}. The algorithm is:

①Figure out that author *C* is the one author *A* quoted most of all among these 1700 authors, the number of times author *C* is cited by author *A* is denoted as C_{maxA}。

②Figure out that author *D* is the one author *B* quoted most of all among these 1700 authors, the number of times author *D* is cited by author *B* is denoted as C_{maxB}。

③Figure out the number of times author *B* is cited by author *A*, denoted as C_{AB}, and the number of times author *A* is cited by author *B*, denoted as S_{BA}.

④Calculate author *A* in author *B*'s citing circle S_{AB}

$$S_{AB} = \frac{C_{BA}}{C_{maxB}} * 10 \quad (4)$$

⑤Calculate author *B* in author *A*'s citing circle S_{BA}

$$S_{BA} = \frac{C_{AB}}{C_{maxA}} * 10 \quad (5)$$

⑥Calculate the score of author *A* and author S_Y

$$S_Y = \frac{1}{2}S_{AB} + \frac{1}{2}S_{BA} \quad (6)$$

This score S_Y' of author *A* and author *B* represents the degree of importance in each other's scholarly circles in the form of quotation. So S_Y' can be split into the degree of importance of author *A* in author *B*'s cited circles which is denoted as S_{AB} and the degree of importance of author *B* in author *A*'s cited circles which is denoted as S_{BA}. The algorithm is:

①Figure out that author *C* is the one who cited author *A* most of all among these 1700 authors, the number of times author *A* is cited by author *C* is denoted as C_{maxA}。

②Figure out that author *D* is the one who cited author *B* most of all among these 1700

authors, the number of times author *B* is cited by author *D* is denoted as C_{maxB}。
③Figure out the number of times author *A* is cited by author *B*, denoted as C_{AB}, and the number of times author *B* is cited by author *A*, denoted as S_{BA}.
④Calculate author *A* in author *B*'s cited circle S_{AB}

$$S_{AB} = \frac{C_{AB}}{C_{maxB}} * 10 \quad (7)$$

⑤Calculate author *B* in author *A*'s cited circle S_{BA}

$$S_{BA} = \frac{C_{BA}}{C_{maxA}} * 10 \quad (8)$$

⑥Calculate the score of author *A* and author S_Y

$$S_Y' = \frac{1}{2}S_{AB} + \frac{1}{2}S_{BA} \quad (9)$$

The importance of academic cooperation relationship and the mutual quotation relationship is the same in the analysis of academic relationship. Therefore, we need to set up weight values of the cooperation relationship and the mutual quotation relationship respectively before calculating the score of academic relationship. Weight value for the cooperation relationship is denoted as *F*, and Weight value for the mutual quotation is denoted as F'. Finally, the score of academic relationship is $S = FS_H + F'S_{HY}$. *F* is set as 0.6 and F' is set as 0.4 in this research. Fig.2 shows some sores of academic relationship.

Author 1	Author 2	S_H	S_{HY}	S
Zuo Tao	Shen Jinxiong	3.2	1.6	2.56
Zhou Xiaorong	Zhu YouYong	0	2.5	1
Zhuang Zhimeng	Jia Jingxian	2.6	1.8	2.28
Zhuang Wen	Huang Danian	0	1.4	0.56
Zhuang Ping	Fei Yanliang	1.3	2.1	1.56
Zhu Chenjian	Zheng Diansher	0	2.3	0.92
Zhu Zengjun	Fen Zhongfu	0	0	0
Zhu YongGui	Liu Jiafu	0.7	1.9	1.18

Fig.2 Partial score of academic relationship

2.2 CLARANS algorithm

CLARANS (Clustering Large Applications based upon Randomized Search) is one of the partition-based clustering algorithms. The aim of the partition-based algorithms is to decompose the set of objects into a set of disjoint clusters where the number of the resulting clusters is predefined by the user. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. It has low efficiency when it scales for large data sets, due to complex iteration process. CLARANS is the most effective portioning method, widely used in large data set. The algorithm is composed of the following steps:
①for *i = 1* to *v* (the number of sampling) , Repeat Steps 2 to 4。
②Draw a sample of *N* (such as 40 + 2*k*) objects randomly from the entire data set and call PAM algorithm to find *k* medoids of the sample.
③For each object O_j in the entire data set, determine *k* medoids which is most similar to O_j.
④Calculate average dissimilarity of the clusters obtained from Step 3. If this value is less than current minimum, use the new value as current minimum and retain the k

medoids found in Step 2 as the best set of medoids obtained so far.
⑤Return to Step 1 to start the next iteration.

2.3 Design of experiments

This study verifies the proposed clustering method for agricultural scientific data based on author's scientific research relationship based on the experimental results.

The experiment set up an experimental group and two control groups. In the experimental group, 5000 data pairs were selected from agricultural scientific data after clustered with the proposed method. Each data pair was selected from the same cluster randomly. 5000 data pairs in control group 1 were selected from agricultural scientific data randomly. 5000 data pairs in control group 2 were selected from agricultural scientific data after clustered with the traditional clustering method. Each data pair was selected from the same cluster randomly.

During the experiment, testers respectively browsed one pair of agricultural scientific data and their metadata item which were selected from the experimental group and two control groups randomly. Testers determined the correlation of the data pair based on their metadata item, and rated the correlation on a scale of zero to ten. A higher score indicated a higher correlation of the data pair.

In ordinary living, there are four categories of personnel who have a potentially high interest in the agricultural scientific data, including agricultural research staff, agricultural technicians, students of agriculture universities and farmers. 50 testers were selected from above four categories of personnel in this experiment, including 18 agricultural researchers (36%), 7 agricultural technicians (14%), 6 students of agriculture universities (38%), and farmers (12%).

In order to ensure the testers can browse scientific data and its metadata quickly and make a comparison in the two datasets of one data pair intuitively, an experimental system was specifically developed, shown in Fig.3. The system showed two datasets of one agricultural scientific data pair to testers at the same time. Testers can browse the specific content and metadata of two datasets through simply mouse operations, and can make a comparison intuitively. Rating area was set up at the bottom of the interface. The scores were filled in here by testers can be automatically saved into the database. In order to minimize the effects applied to datasets experimental data, which were caused by learning effect and accumulation of experience, the experimental datasets showed in system were selected randomly from the experimental group and two control groups each time. So the effects caused by learning effect and accumulation of experience were spread over such three groups.

Fig. 3. Experimental system

2.4 Data and Analysis

2.5

Total 8736 sets of experimental data were obtained. 2927 sets were from experimental group, accounting for 33.51%; 2889 sets were from control group 1, accounting for 33.07%, 2920 sets were from control group 2, accounting for 33.42%. It showed that the experimental data is evenly distributed in the experimental group and two control groups basically. Fig.4-6 show some set for three groups, respectively.

Data set 1	Keywords 1	Data Set 2	Keywords 2
Grassland utilization database	Grassland utilization	Reserve agricultural land utilization resource database	Reserve; Agricultural land utilization; Resource
Feed mineral element and vitamin content database 2014	China; Feed; Mineral element	Feed classification database	Feed classificatio; Animal
nutrient requirements of rabbits database	rabbits; Animal; Requirement	Feed nutritional component Database 1990	China; Feed; nutritional component; Animal
Fish-ball shaping machine database	Shaping machine;Technical parameters	Database on number, distribution, function and status of fishing port	Name of fishing port; Level of fishing port
Cutting machine database	Cutting machine;Technical parameters	Feed mixing machine database	Feed mixing machine;Technical parameters
Dryer Databse	Dryer; Technical parameters	Inland water quality monitoring dabase(upper Yangtze River)	Inland water; Water quality

Fig. 4. Some sets in from experimental group

Data set 1	Keywords 1	Data Set 2	Keywords 2
Biological resource database	Biological resource	Global rescue system	Global rescue system;Technical parameters
Status of livestock and poultry genetic resource database(cattle)	Cattle; Livestock; Animal	Tropical ornamental plant picture datbase	Tropical ornamental plant picture datbase
Database on number, distribution, function and status of fishing port	Name of fishing port; level of fishing port	Fish finder database	Fish finder; Technical parameters
Database of agricultural production and distribution of agricultural products	Agricultural production; Distribution of agricultural products	Database of crop seed storage behavior	Crop seed
China county agricultural industrial structure database	China; County; Agricultural industrial structure	Agricultural regional development and planning database	Agricultural regional development and planning

Fig. 5. Some sets in from control group 1

Data set 1	Keywords 1	Data Set 2	Keywords 2
Grassland policies, regulations and standards database	Grassland science; Policies; Regulations and information	Other land use information database	Other; Land use
Agricultural and rural development strategy database	Agricultural and rural; development strategy	Artemia egg processing equipment	Artemia egg; Processing and technical parameters
Grassland animal Observation Database	Grassland; Animal Observation; field stations	Marine environmental monitoring Fish Database (South China Sea)	Name of vessel; Voyage; Sea area
Fishing law database	Fishing gear; Net name; Fishing boat; Fishing law	Rope making machine	Rope making machine; Technical parameters
Nitrogen grouping of commonly used feed for NRC dairy cows	Feed, NRC; Dairy cows; Scientific data; Animals	Pasture water and fertilizer management database	Name of pasture; Distribution area

Fig. 6. Some sets in from control group 2

Arithmetic mean and standard deviation were calculated in the experimental group, the control group 1 and the control group 2 respectively，as shown in Table 1.

Table 1. Arithmetic mean and standard deviation

	experimental group	control group 1	control group 2
Arithmetic mean	6.84	2.84	6.80
standard deviation	2.01	3.19	2.38

Data of control group 1 were randomly selected from all agricultural scientific data in the share center. The data pairs selected almost had little relevance in most cases. Therefore the arithmetic average of control group was is very low. All sorts of correlation degree were in control group 1 due to random choice, resulting in standard deviation of control group 1 higher. Data pairs of the experimental group and control group 2 were selected from the same cluster after clustering. The correlation of the data pairs selected was quite high, resulting in a significantly increase in their arithmetic mean. The arithmetic mean of the experimental group did not have obvious increase compared with the control group 2, while the standard deviation of the experimental group was obviously superior to that of the control group 2. It indicated that error classification probability of the clustering method used in the experimental group was smaller than that of the control group 2.

3 Conclusion

The traditional clustering method based on the fields of scientific data and the keywords of metadata has the semantic parse and bias problems which are difficult to completely solve. So a part of the scientific data is partitioned into less connected cluster by mistake. The proposed clustering method for agricultural scientific data based on author's scientific research relationship no longer have the semantic parse and bias problems. Hence, the above error has been improved obviously.

There is also a small part of scientific data is partitioned into wrong clusters in the experiment. On the one hand, some of the author had a substantial change in their research fields or upload scientific data for some other people. Thus it leads to worse clustering results. On the other hand, CLARA_NS algorithm sacrifices the stability of

results partly in pursuit of clustering efficiency for a large number of data. So it has a certain impact on the experimental result.

Acknowledgements: Funding for this research was provided by national science and technology basic conditions platform ”The agricultural science data sharing Centre” (2005DKA31800) and technology Innovation Engineering project of CAAS “Research on agricultural cognitive computing and supercomputing” (CAAS-ASTIP-2016-AII).

References

1. Xiuhui Wang, & Xubiao Yin. A Study On Scientific Data Clustering Based on Improved FIHC, Journal of Shanxi Datong University (Natural Science). 30, PP.4-7(2014).
2. Ingwersen, P., & Jarvelin, K.. The Turn Integration of Information Seeking and Retrieval in Context (2005).
3. Yunxia Zhu. An Empirical Study on Correlation Degree Model of Scholars Based on Co Authorship and Reference relation Analysis, Library Science and Informatics. 22, PP.97-103, (2015).
4. Jiang Wan, Dingfeng Wu. Design and Improvement of Single Sign-on Technology for Agriculture Information Services. Computer Technology and Development. 26(5), 191-196(2016).
5. Suhui Wu. The Improvement of Academic Retrieval System Based on Citation Analysis (2012).
6. Ning, Cai, D. Chen, and M. J. Khan. A Novel Clustering Method Based on Quasi-Consensus Motions of Dynamical Multiagent Systems (2017).
7. Dileep Kumar Yadav. Comparative analysis of clustering Current Research, 7, , 18361-18364 (2015).
8. Kamalpreet Kaur Jassar and Kanwalvir Singh Dhindsa. Article: Comparative Study and Performance Analysis of Clustering Algorithms. IJCA Proceedings on International Conference on ICT for Healthcare ICTHC 2015(1):1-6, (2016).
9. Wang, Wei, et al. Relationship between Academic Cooperation and Research Output Based on Academic Journal Papers: A Research Taking the Library and Information Field as an Example. Journal of Intelligence (2017).
10. Yang, Ruixian, and M. Zhang. A Research Review on Academic Relationship of Authors. Library & Information Service (2016).
11. Zhang, Shan Hong, et al. Study on the Relationship between College Students' Academic Activities and Academic Performance. Value Engineering (2018).
12. Wang, Xiu Hui, X. B. Yin, and B. O. Wen-Yan. Improved Scientific Data Clustering Algorithm Based on FIHC. Journal of Shanxi Datong University (2014).
13. B. E. Husic and V. S. Pande, Ward clustering improves crossvalidated Markov state models of protein folding, J. Chem. Theory Comput. 13, 963−967 (2017).
14. Brooke E. Husic, Kristy L. Schlueter-Kuck, John O. Dabiri. Amplifying state dissimilarity leads to robust and interpretable clustering of scientific data. Machine Learning(2018)