Research on Tang Poetry Collection from the Perspective of Big Data

Wang Fangfang
Zhengzhou University of Industrial Technology, Zhengzhou, Henan, 451100, China

Abstract

Since entering the 21st century, computer technology has been popularized and developed rapidly. Using big data to study humanities has gradually become a research trend. More and more scholars begin to pay attention to and play the role of big data in the study of classical literature. With the rise and application of academic map based on big data, new research paradigm will promote and expand the research field of Tang poetry collection. As a result, the Electronic Full-text archives of Quan Tang Shi have been circulated in the world for a long time. So that scholars can apply the relevant archives, gradually introduce the method of Digital Humanities, and carry out various data statistics and analysis from various angles. Based on the big data technology, this paper gradually counts all kinds of data of Tang poetry, in order to find out the issues worthy of exploration. At the same time, through the interpretation of a series of data, it echoes the research results of traditional Tang poetry. Experimental research shows that with the help of relevant big data, it can expand a new vision of Tang poetry research

Keywords*: Big data, data mining, Tang poetry, data fusion*

I. Introduction

With the popularization and continuous development of computer technology, the research of humanities with big data has gradually become a research trend. As early as the 1990s, Mr. Zhu Zejie prepared to compile the whole Tang poetry, and began to use computers for compilation and research [1]. Combined with the compilation of the whole Tang poetry, he wrote and published several related papers. After entering the 21st century, on the basis of the popularization of ancient books digitization and the continuous improvement of big data technology, the databases with electronic and full-text retrieval as the main functions can not meet the needs of research. Composite database with more functions has become the common demand of database designers and users. The research and application of big data began to move from digitalization to digitalization [2-4].

In recent years, the rise of academic map is based on data. As early as 2012, Professor Wang Zhaopeng put forward the idea of building a digital map platform for Chinese literature [5], and in the same year, he was approved as the national social science major bidding project "construction of local information platform of Tang and song literature chronology Department". In March 2017, the "chronological map of literature of Tang and Song Dynasties" produced by Professor Wang Zhaopeng was launched on soyun [6] (https://souyun.cn /poetlifemap.aspx)It is not only highly praised by the academic circles, but also attracted the attention of many poetry lovers. On March 19, 2018, the academic map publishing platform jointly built by Harvard University and Zhejiang University [7] (http://amap. zju.edu.cn) officially launched, it has released more than 300 academic maps, more than 600 layers and more than 400000 pieces of data (as of November 25, 2018).

The rise of big data and academic map provides a new perspective for the visualization of classical literature research. With the help of database and software, we can realize the visualization of the writer's track and activity place, the geographical distribution of the writer's social relations and the social relations of the characters. These visualizations are intuitive and refreshing, and will play an excellent auxiliary role in the research and teaching of Chinese classical literature. As far as the research field of Tang poetry anthology is concerned, the Tang Dynasty literature resources published by the academic map publishing platform can be used for reference. The new research paradigm embodied in the academic map has certain enlightening significance [8].

II. Inspiration of big data academic map to the new paradigm of classical literature and Tang Poetry Research

For a long time, the study of classical literature has focused on qualitative analysis, emphasizing the personalized interpretation of the text and writers, but the quantitative analysis is insufficient. The traditional textual research is often to research a specific problem, and the conclusion is not the same as the data today. With the development of science and technology, quantitative analysis of text using big data has become a new research method, and has produced a number of influential academic achievements. On the other hand, the academic map marked by the "chronological map of Tang and song literature" has provided a new paradigm for the study of classical literature. This paper holds that the concept of "integration of time and space" and "panoramic view of literary history" embodied in the chronological map of Tang and song literature are of great significance to the study of classical literature. First of all, the characteristics of the academic map in time and space are connected with each other, which solves the problem of time and place separation in the field of literary research, and plays an important role in compiling the chronology of works and writers.

The study of classical literature has accumulated fruitful results in the field of chronology of works and chronology of writers. The study of regional literature with space as the main line is often carried out from the perspectives of schools, families and teachers. To a certain extent, there is the problem of time and place separation. The advantage of academic map lies in the combination of time and space. When querying a writer or a work, users can obtain both time and space information at the same time, and can switch time and space according to their own needs. This makes up for the deficiency of the past single research focusing on time or space. The introduction of the concept of space (region) into the compilation of works and writers' chronicles not only breaks the traditional way of chronology, but also provides a richer perspective for the study of works and writers. It is helpful for readers to better understand and grasp the connotation of the works by investigating the works with time and place to the specific space-time environment and restoring the writing background at that time. In the investigation of the writer's life story, in addition to combing with the year as the clue, we can also locate and search in space according to the author's activities and creations in different places, simulate and reproduce the writer's life track, which is helpful for readers to understand and study the writer's personality. Secondly, the application of academic map presents the characteristics of "panoramic view of literary history", which solves the problem of scattered achievements in the field of literary research.

The research object of chronological map of literature in Tang and Song Dynasties is not a certain writer or a certain work, but famous writers and their works with works handed down throughout the Tang and Song dynasties. Through the map, it presents the life track and works of Tang and song writers in a panoramic way. The traditional way of academic research is to collect and count the research results by hand. In the face of massive information, it is often lack of clue or missed, and the efficiency is not high. The chronological map of Tang and song literature shows the research results in the form of data. Users can search and save relevant research data according to their own needs, and establish their own research database, which is not only efficient, but also can be adjusted and saved at any time. This lays a foundation for quantitative analysis based on big data. There is no contradiction between qualitative analysis and quantitative analysis in classical literature research. Through quantitative analysis, we can detect and identify whether the qualitative analysis is accurate or reasonable.

There are many collections of Tang poetry. More than ten years ago, when he was writing a summary of Tang Poetry Research for history of Tang Dynasty · records of classics · records of literature and art · poems and CI Pian, he proposed that there should be more than 2000 kinds of poems in the world according to the description of "records of art and literature of Tang historical manuscripts", "supplement to records of art and literature of Tang historical manuscripts" and "collection of records of art and literature of Tang historical manuscripts". This figure was inherited by Xia Yong's general collection of Tang poetry. After several years of further investigation, in addition to the completion of the "general catalogue of works of the Tang Dynasty" project presided over by Professor Du zexun in recent years, and the relevant results of the national census of ancient books have gradually come out, I am afraid that the number of 2000 should be revised to 5000 or even more than 10000. But no matter 2000, 5000 or 10000, we can imagine that this is an extremely rich literary and cultural heritage, and also a colorful, complex and wonderful academic world. However, such a rich and profound literary and cultural heritage has not attracted enough attention for a long time, which makes the collection of Tang poetry a weak link in the study of Tang poetry and even the whole study of Tang literature and culture.

III. Semantic clustering model of Tang poetry based on feature word clustering

3.1 Research on nonlinear feature extraction of feature words by clustering

Research shows that the classification information carried by feature words is closely related to the training set provided. Because the semantics of feature words in different contexts are different, and most languages have polysemy, the same feature word in different training sets will not contain exactly the same classification information, or even very different. Combined with previous studies and considering the property of feature words, we represent feature words in a class information vector space. Each dimension of the vector space represents each class provided by the training set, and the weight of each dimension represents the measurement of each class classification information. As shown in Figure 1, the N-dimensional space represents N different categories provided by the training set, and the similarity of different feature word vectors is calculated by some distance formula (such as Euclidean distance, cosine angle, etc.) in the category information vector space.

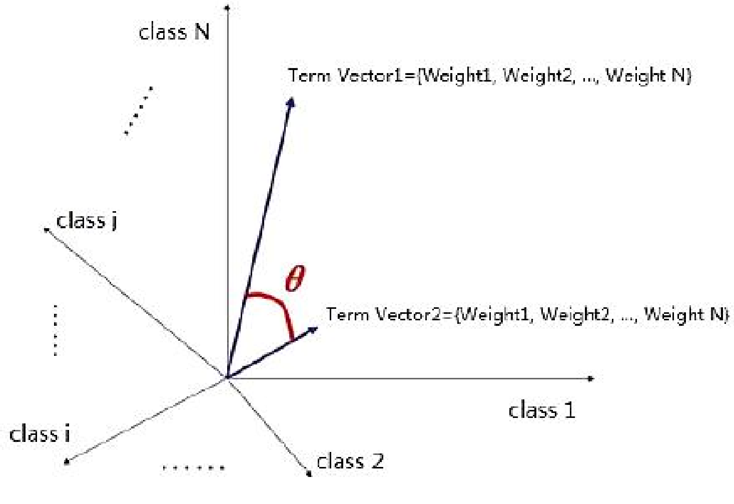
class N
Term Vector1={Weight1, Weight2, ..., Weight N}
class j
θ
Term Vector2={Weight1, Weight2, ..., Weight N}
class 1
class i
class 2

Figure 1 Class information vector space

Generally speaking, the main clustering methods are divided into the following categories:

1. partition method: divide n data into k groups, and each data can only belong to one group. the representative methods are k-means algorithm and k-center algorithm.

2. Hierarchical method: the given data is decomposed hierarchically, and the representative methods are condensation algorithm AGNES and splitting algorithm DIANA；;

3. Density-based method: clustering based on data "domain" density. The representative methods are DBSCAN algorithm and its extended algorithm OPTICS；;

4. Grid-based method: clustering in the grid structure of data space quantization, the representative method is STING algorithm;

5. Model-based method: clustering by finding the best fit of data to a given model, the representative method is EM algorithm.

Among them, the most influential methods in the field of natural language processing are the first classification method and the second hierarchical method. Other clustering methods are rarely used in the field of natural language processing, so this chapter will not repeat them.

The most commonly used method in feature word aggregation is the first class partition method, which is based on the calculation of the distance between feature words. In order to achieve the global optimization, we usually need to enumerate all possible partitions, and the number of possible partitions is given by people. It is a representative class in clustering method. In contrast, the feature word vector represented in the class information vector space proposed in this chapter has the characteristics of low dimension (the same as the number of training set categories), large weight difference, and difficult to determine the number of clustering partition. The partition method that needs to be manually assigned partition score has great uncertainty of clustering results, so the partition method is not suitable for feature word vector clustering.

Hierarchical method can cluster the original data according to the principle of large similarity within clusters and small similarity between clusters without guidance vector. Google has successfully applied this method to news classification for a long time. Figure 2 shows the general steps of two hierarchical clustering methods: aggregation and splitting. Because hierarchical clustering method determines classification level artificially, taking AGNES shown in Figure 2 as an example, if clustering is stopped at Step2, the clustering results are ab, C and de, and if it is stopped at Step3, the results are ab and cde. The user's control of clustering level indirectly achieves the granularity control of classification results, which is very suitable for clustering data with hierarchical gradient.

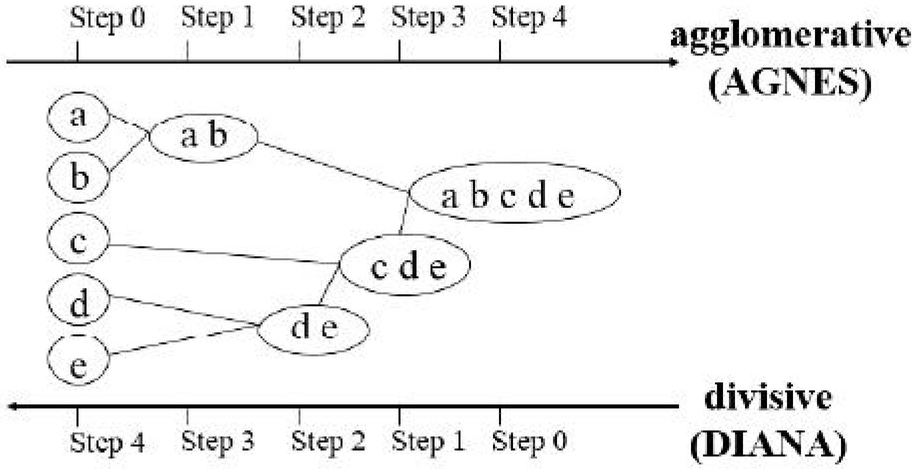

Figure 2 Hierarchical clustering

3.2Constructing semantic cluster model

In this section, we propose a semantic cluster model based on feature word clustering, which constructs a semantic cluster by clustering the words with the same or related semantic features in the training set. To a certain extent, the correlation between feature words is preserved. At the same time, on the premise of retaining the classification information as much as possible, the dimension of the vector space model based on feature words is greatly reduced. The definition of semantic cluster is as follows.

Definition: semantic cluster is a set of feature words. The feature words in the cluster are all from the training set. All the feature words in the same semantic cluster are related to the same one or more training set categories. If a semantic cluster is related to n categories, it is called n-order semantic cluster.

Figure 3 shows the second-order semantic cluster model. The less the order, the clearer the classification information contained in the semantic cluster. The semantic cluster above the second-order points to multiple classifications. Most feature words belong to the semantic cluster above the second-order, and most real words except function words belong to the semantic cluster below the fifth order.

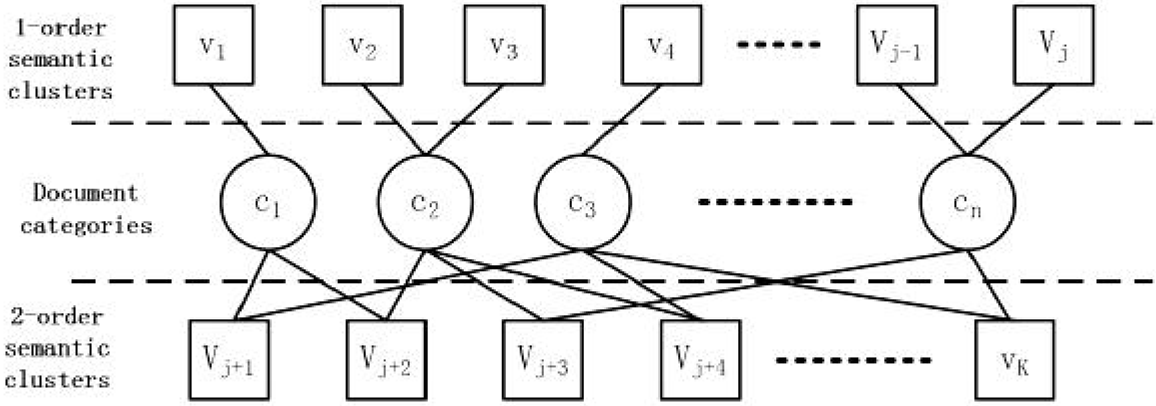

Figure 3 Second order cluster model

Given the training set Corpus={c1,c2,...,ck,...,cN}, where ck is the k-th document subset, the feature word set extracted by the training set corpus is T={t1,t2,...,ti,...,tM}, where the feature word ti is an N-dimensional vector ti = {wi1, wi2, ...,win}.

Figure 4 shows the structure of the traditional text classification system based on machine learning. In the system, the offline training text classifier part and the online classification part are clearly separated. The text model module and the decider are not marked separately in the figure, which are represented by the text classifier module. The feedback operation is performed by modifying the copy of the running text classifier offline. The key of the whole system is the trained text classifier module. In practical application, once the server storing the module needs maintenance or fails, the whole classification system will not work. When the module needs to add new data content or update, it will encounter the same problem.

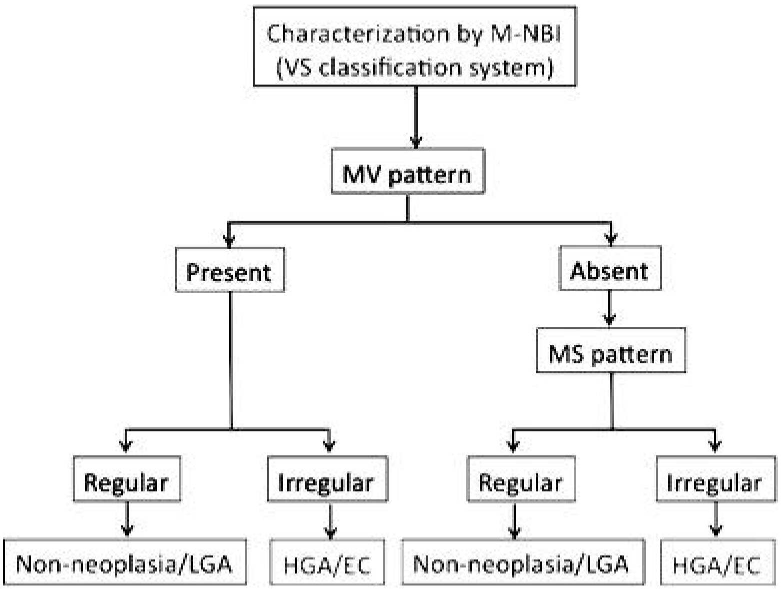

Figure 4 Structure of text classification system based on machine learning

Fig. 5 shows the structure of the SC-HMM text classification system provided in this paper. Different from the text classification system based on machine learning shown in fig. 4, firstly, the semantic cluster model is shared with the online end after the offline end is established for the serialization of text data. in addition, the HMM text model module is shown as a whole in the figure. However, in fact, all kinds of HMM text models are independent of each other, without inclusion or dependency. A single type of HMM text model needs offline training, but other types of HMM text models can work normally online, which has no impact on the overall operation of the system. In the part of correcting classifier weight by feedback of classification results, it is necessary to add documents and classification results to semantic cluster model, and retrain the corresponding HMM text model after serialization.

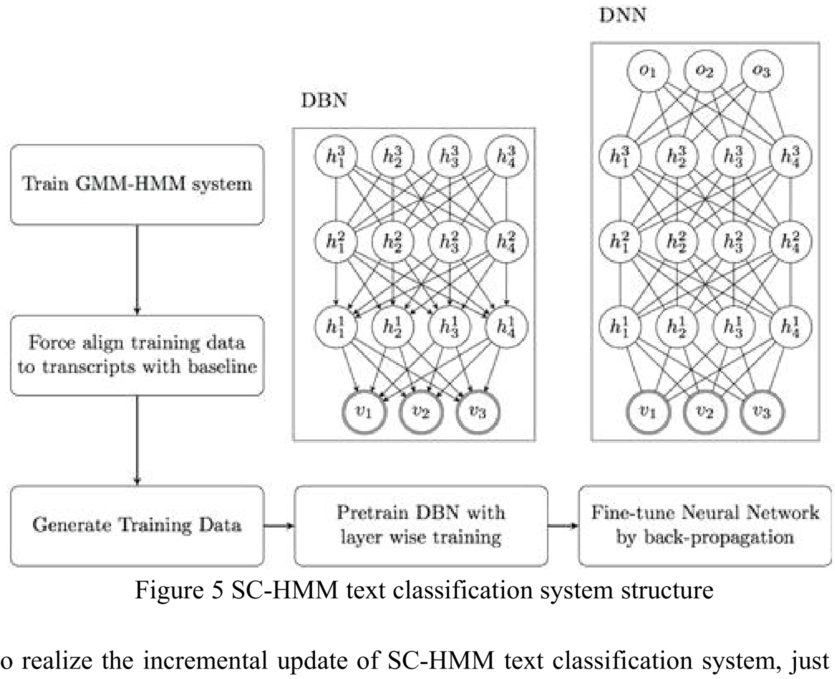
DNN
DBN
Train GMM-HMM system
Force align training data to transcripts with baseline
Generate Training Data
Pretrain DBN with layer wise training
Fine-tune Neural Network by back-propagation
o_1 o_2 o_3
h_1^3 h_2^3 h_3^3 h_4^3
h_1^2 h_2^2 h_3^2 h_4^2
h_1^1 h_2^1 h_3^1 h_4^1
v_1 v_2 v_3

Figure 5 SC-HMM text classification system structure

It is very simple to realize the incremental update of SC-HMM text classification system, just add new categories and corresponding documents to the semantic cluster model update in the off-line training part, and then train the HMM text model of this category separately and add it to the module. At this time, the normal operation of the online classification part is not affected at all, only the evaluation of HMM text model of new category is needed in the probability evaluation part, and the classification number of new category is added in the judgment classification part, and all the work is completed. Similarly, the online update of the system only needs to train some HMM text models with data update and update the corresponding models in the same way.

IV. The future research direction of Tang Poetry

In recent years, with the publication of dissertations and the establishment of national projects, the research on Tang poetry has attracted more attention in the academic circles. In the future, the research on the collection of Tang poetry should start from two aspects: the construction of basic materials and the research of special topics.

There are six directions for the construction of basic materials: first, to investigate the family background of the extant collection of Tang poetry and to list the bibliography of extant collection of Tang poetry; Secondly, he wrote a summary for the extant collection of Tang poetry and compiled a summary of the general catalogue of Tang poetry; Thirdly, the systematic investigation of the lost collection of Tang poetry; Fourth, compile the preface and postscript collection of Tang poetry; Fifthly, the relevant literature review up to 1949 is compiled to provide references for the study of Tang poetry anthology; Sixth, to organize and publish a series of Tang poetry. The above six research directions are closely related to the construction of the research database of Tang poetry collection. The construction and development of Tang poetry research database is not only to transform the research results into data resources synchronously, but also to provide more diversified and personalized services for researchers and preliminarily realize the human-computer interaction under artificial intelligence. At present, the construction of geographic information database of ancient literature includes the distribution of native place of group writers, the distribution of works, the distribution of writing places and the places involved in the content, the trace map of first-class and second-class writers, the location of writers' activities in time series, etc.

This paper thinks that the construction of research database of Tang poetry anthology can refer to his suggestions. First, the geographic information database and academic map should be established for the regional Tang poetry collection, with the province (including the administrative divisions of Tang Dynasty and today's administrative divisions) as the unit. Through the map, we can see the distribution of the number of Tang poetry anthology editors in each province. Secondly, for the collection of Tang poetry, we should establish a database with longitude and latitude, and locate it on the map. Third, establish the personal information database of Tang poetry anthology editors. Fourth, the pictures and related information of stone inscriptions, inscriptions, manuscripts and other objects related to the collection of Tang poetry are imported into the database. These four tasks can be carried out step by step in the specific research work. And it can cooperate with academic map publishing platform to share data resources.

V. Conclusion

The inspiration of big data academic map to the new paradigm of classical literature and Tang poetry research, Tang poetry research is gradually becoming a new academic growth point. As an important field of Tang poetry research, the study of the whole Tang poetry will attract more and more scholars to join in it. With the rapid development of science and technology, big data technology is applied to the study of classical literature. Through the research and development of academic map and the construction of multi-functional database, we can create a new research paradigm and promote and expand the research field of Tang poetry.

References

[1] Dong Wanyue, Bai Ruhai, Chen Xiaotong. Spatial Analysis of Health Human Resource Allocation Level in China. China Health Policy Research, 2018, 12 (3)

[2] Qiu Ligu, Si Yafei, Zhang Xuewen. Study on Human Resource Allocation of Different Types of Community Health Service Centers. China Health Management, 2019, 036 (002): 93-96

[3] Wang Zhigui. on the Standard of Rational Allocation on Human Resources. Journal of Fujian Normal University: Philosophy and Social Sciences

[4] Su Ning, Peng Ying Chun, Wang Ya Dong. Research on Human Resource Allocation of Community General Practice Team Based on Work Analysis. Chinese Journal of General Practice, 2010 (28): 3147-3149

[5] Qian Weining, Zhou Aoying. Analysis of Existing Clustering Algorithms from Multiple Perspectives. Acta Software Sinica, 2002 (08): 1382-1394

[6] Xie Kunwu, Bi Xiaoling, Ye Bin, Xie Kunwu, Bi Xiaoling. High Dimensional Data Clustering Algorithm Based on Unit Region. Computer Research and Development, 2007, 44 (9): 1618-1623

[7] Du Lanying, Zhang Shanjin. Human Resource Allocation and Labor Market Construction in China. Business Research, 2006 (13): 96-98

[8] Zhang Yali, Cai Junping, Cai Jue. Investigation and Analysis of Nursing Human Resource Allocation and Demand in Traditional Chinese Medicine Hospitals. Chinese Journal of Nursing, 2007,42 (011): 1028-1030

[9] Tian Xiuyun. Ethical Thinking of Human Resource Allocation and Management. Morality and Civilization, 2001, 000 (002): 33-36

[10] Wang Jinxiu, Huang Hongfa. Discussion on Human Resource Allocation and Management Mode in Modern Enterprises. Jiangxi Social Sciences, 2003 (12): 159-162