Application of Association Rule Mining in Chinese Part-of-Speech Data

Guijie Wu*[1], Qingqing Lin[2]
[1]Beijing Polytechnic of Information Technology, China
[2] Huawei Cloud Computing Technology Co., Ltd, China

ABSTRACT

The development of the information age has brought people rich and diverse data information, changed people's life and work mode, and raised the requirements of information mining and application. After entering the era of big data, in the face of the ever-increasing network information and data, how to mine valuable content from it and apply it into practice is the main topic of comprehensive exploration by scholars in various fields. Data mining and knowledge discovery technologies are the theoretical technologies generated from it, which have been comprehensively promoted in practice and development, and show strong vitality. It is one of the important fields of database research. On the basis of understanding the current situation of Chinese text data mining, this paper constructs a data mining model with data warehouse as the core according to the basic concepts, implementation process and knowledge types of data mining, clarifies the association rule algorithm applied in Chinese parts-of-speech data, and deeply discusses the structure and application function of Chinese parts-of-speech data and its association mining system. Finally, more valuable content is used in data analysis.

Key words: Association rules; Data mining; Chinese magnetic data; Text mining

1. INTRODUCTION

As a basic research topic in the field of Chinese information processing, automatic tagging of Chinese parts of speech can be used for further study of Chinese syntax and information retrieval. Therefore, relevant research topics are very important. The current world is an information age, the Internet and media have greatly enriched our lives, the media, scientific research and other aspects of the data information is more and more. In the face of so much information, it is the main direction of scholars in various fields to analyze the important content hidden behind the information, find the content hidden between the information, and make effective use of it. With the continuous development of computer technology, data management technology has gone through three stages: manual management, file system and database system. The database data model has gradually developed from the early hierarchical model and network model to the relational model. Nowadays, there are four kinds of common relational data: OB2, Oracle, SQL, Sever Sybase, which can be used to store a large amount of data information and efficiently realize basic functions such as data entry, data query, and data statistics. The development of relational database system has effectively improved the operation and management efficiency of enterprises, so most enterprises choose to use management information system to deal with daily work, and these management systems have accumulated rich information for enterprises. From the perspective of practical application, relational database system will use a single data organization mode to work, and its advantage lies in connecting business processing. It is difficult to find hidden data relations and data rules relying only on traditional database system, and it is impossible to predict the future development trend based on existing data. Therefore, some enterprises begin to use information to provide support for management decisions. Therefore, the environment requirements of traditional database system are changed. With the continuous development of modern science and technology, a new database technology has rapidly developed in the field of information, which is called data warehouse, which can meet the diversified data processing needs and can be used together with data mining technology to complete the management and analysis of massive data information.

As a new branch in the field of data mining, text data mining, which has received public attention in recent years and developed rapidly, is closely related to advanced theoretical methods such as machine learning, data statistics and pattern

[1]*wugj@bitc.edu.cn

The International Conference Optoelectronic Information and Optical Engineering (OIOE2024), edited by
Yang Yue, Lu Leng, Proc. of SPIE Vol. 13513, 135133F · © 2025 SPIE · 0277-786X · © The Authors.

analysis. Text data mining is divided into data mining with single file as the core and data mining with document set as the core. The former does not include other documents when analyzing documents, and the specific mining directions include automatic text summary, document knowledge summary discovery, information extraction, etc. The latter will extract large-scale document data, including not only automatic document summary and document summary, but also document classification, text clustering, similarity analysis and personalized text filtering. From the perspective of practical application, text data mining mainly uses feature information extraction, cluster analysis and other methods to classify text, and has achieved excellent results in informatics and book information retrieval, etc. A small number of people can use linguistic grammar structure and knowledge to analyze text content, but the progress is slow. This paper discusses the method and theory of Chinese magnetic data acquisition from the perspective of data mining, and finds more information hidden in the data under the condition of meeting the user rules, so as to improve the efficiency and quality of mining.

2. METHOD

2.1 Data Mining

Data mining refers to the discovery of hidden, regular, unknown, potentially useful and ultimately understandable information and knowledge from a large number of incomplete, noisy, fuzzy and random practical application data. From the perspective of practical application, data mining is an iterative process, and different application requirements and different data sources will change the data mining steps. According to the data mining flow chart shown in Figure 1 below, CRISP-DM divides the whole mining process into six stages, in which business understanding refers to the understanding of enterprise operation, business process and industry background, and data understanding refers to the understanding of the current enterprise application system. Data preparation refers to obtaining a sample data subset related to exploration problems from a large amount of enterprise data. Modeling is to understand and select practical mining models based on business problems, and to evaluate and obtain conclusions in inspection mining. Only when the expected results are achieved can the conclusions be released.

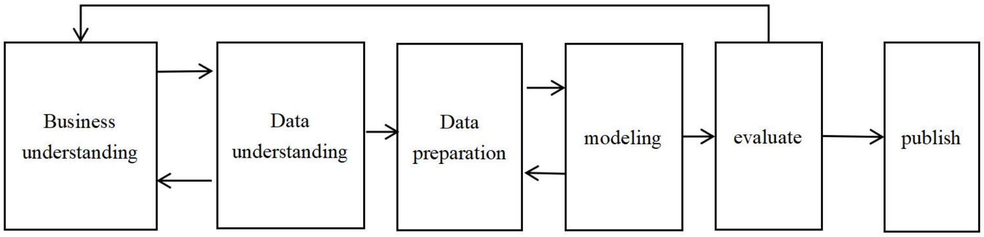

Figure 1. Flowchart of CRISP-DM mining.

2.2 Data mining model based on data warehouse

When dealing with the business system with huge amount of data and different analysis angles, it is necessary to build a data mining system with data warehouse as the core. Combined with the data warehouse structure diagram shown in Figure 2 below, we can see that it should be a solution for the user, not a product. Nowadays, most data warehouse systems are based on relational database systems. In order to extract the acquired source data and form a comprehensive data form that can be used for decision analysis, the data warehouse is divided into four levels: data source, data extraction and transformation loading, target data warehouse, data access and analysis. The purpose of building a data warehouse is to obtain valuable regular knowledge for decision management activities from a large amount of data, which can be operated by multi-dimensional analysis, complex query, connected transaction analysis and data mining technology. Data mining is an important tool to discover useful knowledge in mass. The data mining model based on data warehouse is shown in Figure 3 below:

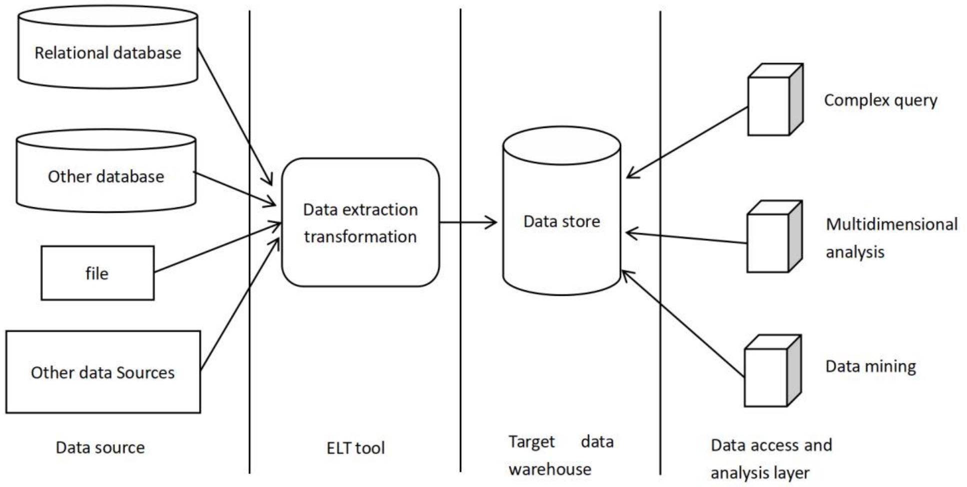

Figure 2. Architecture diagram of data warehouse.

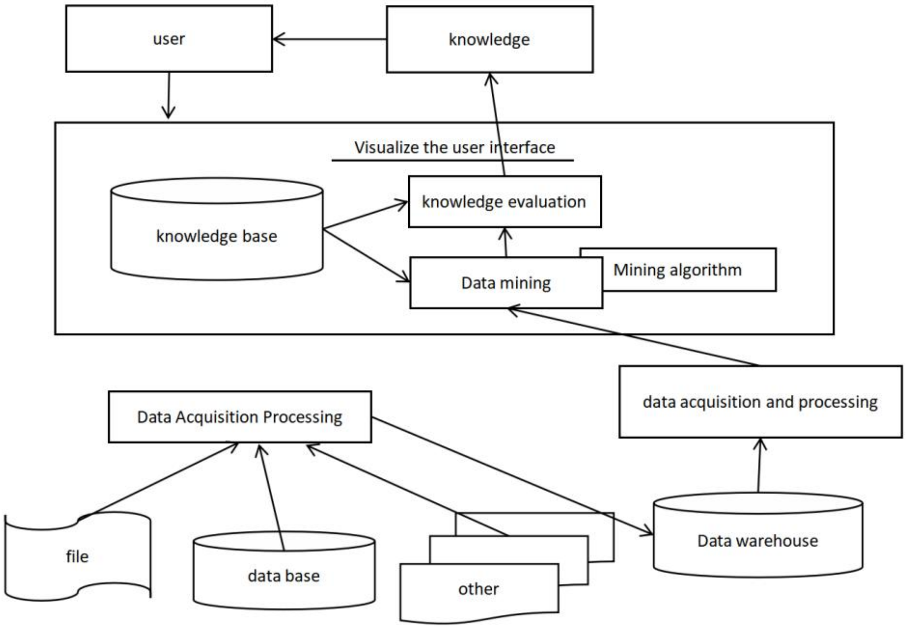

Figure 3. Data mining model diagram based on data warehouse.

According to the above analysis, the overall model is divided into two parts, one is the index data warehouse construction model, the other is the index data mining model, both of which include data collection and processing, and are important links in data warehouse and data mining.

2.3 Association rule algorithm

Association rules refer to the existence of correlation or interrelation between various data in a large amount of data. Association rule data mining is the search for connections between data lines in a given data set. Association rules will describe the degree of interphase relationship between a set of data sets. As one of the important research topics of data mining, the corresponding technical algorithms have been highly valued by the academic community and widely used in enterprises, scientific research, business and other aspects. For large databases, the application of association rules mining algorithm can eventually get thousands of rules, in these rules, some of the rules belong to false rules, so in practical application, algorithm optimization and improvement should be carried out according to demand. The design of association rule mining system based on Chinese parts of speech data must take into account that text data is a data body with limited structure or even no structure, and the text format is likely to have differences in forms such as paragraphs, indents, text, graphics and tables. Therefore, it is necessary to preprocess data information and extract source data that can represent its features. Then efficiently process the already structured source data. Combined with the model architecture analysis shown in Figure 4 below, we can see that the overall system operation is divided into the following steps:

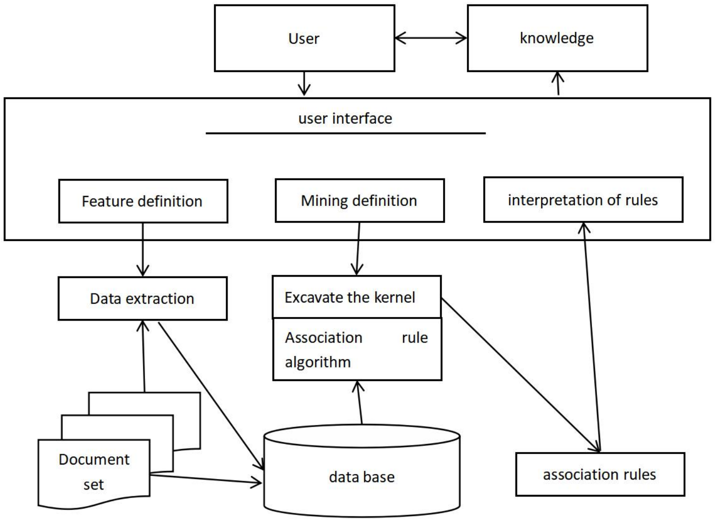

Figure 4. Structure diagram of association rules mining system based on Chinese part-of-speech data.

According to the above analysis, practical application modules are divided into the following points: First, determine the target sample. According to the user's choice to determine the mining target text sample, using the data extraction module to extract the text feature value; Second, data extraction. After obtaining the user-specified set of plain text files, the feature number in the text is extracted according to the specified minimum statistical technique of feature value. The extraction method without background knowledge can be used to analyze the frequency of occurrence of two-word group and three-word phrase in the text. If the frequency exceeds the specified minimum count, the phrase will be regarded as the eigenvalue of the text. Third, feature value cleaning. Characteristic values obtained from data extraction to be cleaned. The extraction of text files is read out according to the position in the text. Some control characters without natural semantics appear in some positions, which will be read into and form meaningless garbled codes. If these garbled codes exceed the minimum statistical value, they will also be added to the feature value, so they must be cleaned to ensure the accuracy and consistency of the data stored in the database. Fourth, the eigenvalues are stored. Add the feature value after data cleaning to the database, and the relevant attribute values of the text file, such as file name, file path, creation time, storage time, etc.; Fifth, association rules mining. After the feature value is loaded into the database, each text file corresponds to a feature reference vector, which is similar to the transaction identification number and transaction record in the commodity transaction database. Weighted association rule algorithm can be used to mine the feature value vector and obtain the association rules. Sixth, using TextMiner, the association rules mined by the system can be used for unknown knowledge discovery and text content retrieval. Other extended applications include intra-text office, automatic text keyword extraction, automatic text summary, etc.

3. RESULT ANALYSIS

In the TextMiner text mining system, strong association rules are mined using the mining kernel composed of association rules data mining algorithm, and then the mined rules are interpreted by the rule interpreter, and finally valuable knowledge is formed and presented to users. The specific design process is shown in Figure 5 below:

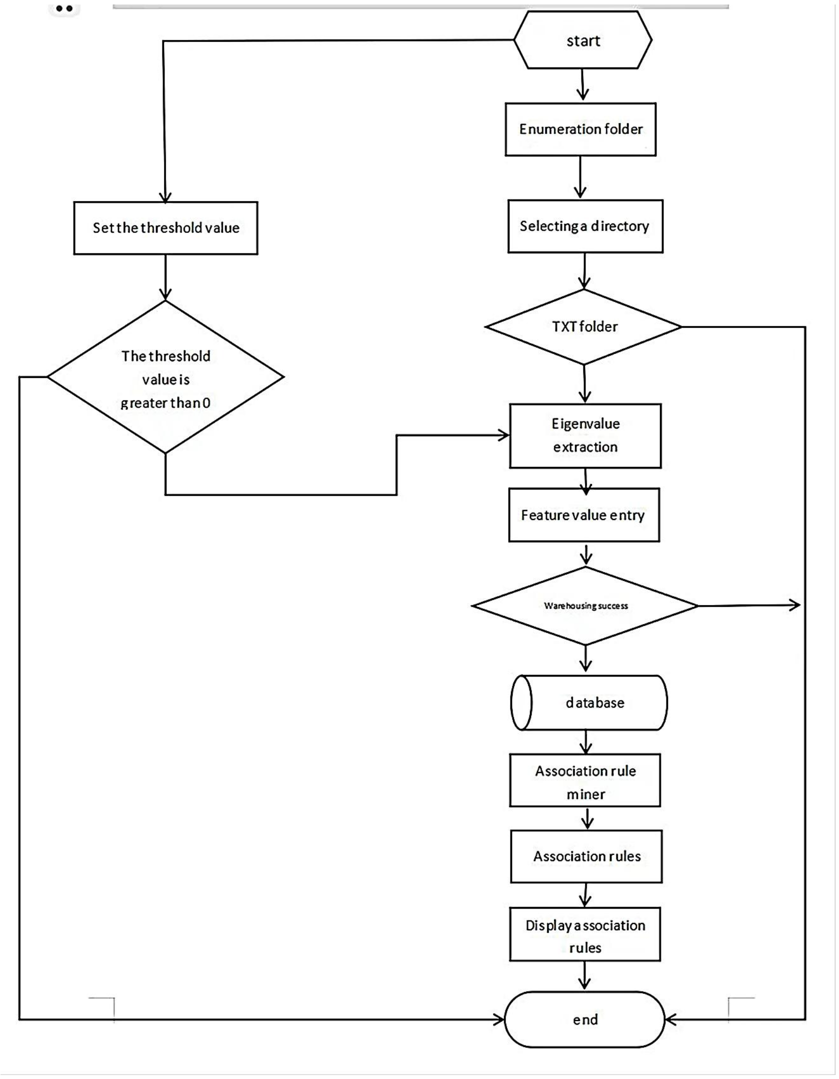

Figure 5. Flowchart of the program design.

From the perspective of practical application, this system makes clear the relationship between data mining technology and data warehouse, deeply discusses the association rules data mining algorithm, and clarifies the application value of various algorithms. At present, information retrieval has made excellent achievements in text structure analysis and text classification, and researchers should continue to explore related technologies of text content mining in the future. Since text content mining has just started and has a broad practical application prospect, it has become an important topic in information retrieval and intelligence analysis. The application of text content mining research results in search technology can provide people with more accurate search results, so there will be more and more research content on the design of Chinese partof speech data and its association rules mining system. The processing efficiency and quality of textual data will be improved.

4. CONCLUSION

To sum up, data mining, as one of the important contents of the development and innovation of database technology, can mine valuable content from a large amount of data information, which is convenient for people to efficiently use data resources in the new era. Therefore, in the face of a large number of Chinese magnetic data information emerging in the new era, it is necessary to describe the interesting links between a given data set, perfect mining all association rules, effectively extract text feature values, and finally provide people with more accurate search results to ensure the accuracy and representativeness of feature values. Only in this way can data mining efficiency be improved. Ensure data application value.

REFERENCE

[1] Jiyao Lei. Application research of data mining algorithm based on association rules in the field of e-commerce [J]. Information and Computer, 2023, 35(16):73-75. (in Chinese)

[2] Xingyun Ji, Jianhua Fan. Study on Drug administration of Longan meat prescription in CNKI based on Python language [J]. Wisdom and Health, 2023, 9(15):153-160.

[3] Liu Yang, Qingwen Xu, Rujia Huang, et al. Application of association rules in traditional Chinese medicine data mining [J]. Information of Traditional Chinese Medicine, 2022, 39(12):35-40.

[4] Yuezhou Zhao, Haoran Li, Bohua Chen, et al. Study on drug use of homologous Chinese medicines in treating cough after cold based on data mining [J]. Inner Mongolia Traditional Chinese Medicine, 2023, 42(3):162-166.

[5] Qiaoli Liang, Menglong Zou, Xiaoyan Huang. Study on the law of TCM syndrome differentiation and treatment of functional dyspepsia based on data mining [J]. Guangxi Medicine, 2023, 45(22):2749-2755. (in Chinese)

[6] Feiyan Mu, Shuxun Shi, Qinqin Liu,et al. Medication regularity and mechanism of traditional Chinese medicine for hemorrhoidal postoperative complications: based on literature data mining and network pharmacology [J]. Asian and Pacific Traditional Medicine, 2024, 20(1):154-162.

[7] Qing Xia , Qingyang Ou, Dan Shen,et al. Application and prospect of establishing association rules for catheter-related infections based on data mining technology [J]. Journal of General Nursing, 2023, 21(23):3214-3217.

[8] [8] Xinyu Zhu, Yao Zhu, Ming Lu , et al. Study on the application of activating blood circulation and removing blood stasis in multiple myeloma based on data mining and 13 pathologic mechanisms [J]. New Chinese Medicine, 2023, 55(7):155-159.

[9] Nan Guo, Bo Yang, Hongru Liu,et al. Study on medication regularity of TCM external treatment for pruritus associated with chronic kidney disease based on data mining [J]. Chinese Information Journal of Traditional Chinese Medicine, 2024, 31(2):21-26.

[10] Liqin Xu, Wenzhe Han, Huali Luo. Study on the rule of external use of traditional Chinese medicine for atrophic vaginitis based on data mining [J]. Shanxi Traditional Chinese Medicine, 2024, 40(1):49-52.

[11] Zhihong Zhan, Guohua Dai. Analysis of relevant literature in the field of TCM drug use rule data mining based on Citespace [J]. Inner Mongolia Traditional Chinese Medicine, 2023, 42(3):132-137.

[12] Haizhen Guo, Han Wu , Meikang Mo,et al. Visualization data mining analysis of electroacupuncture therapy for vascular cognitive impairment [J]. Journal of Guangzhou University of Traditional Chinese Medicine, 2024, 41(1):161-168.

[13] Xianyu Wu , Longmei Yan, Yaxuan Xing.et al. Study on medication rule of Qinggong Medical Record Integration in treating insomnia based on data mining technology [J]. Journal of Cardio-Cerebrovascular Diseases of Integrated Traditional Chinese and Western Medicine, 2024, 22(2):239-243.

[14] Shuo Li, Peipei Zhao, Bin Yuan. Study on the rule and mechanism of Chinese medicine in treating adenoid hypertrophy in children based on data mining and network pharmacology [J]. Medical Information, 2024, 37(2):55-63.

[15] Xicuo Ye, Jie Renqing Duo, Sang Geng,et al. Study on drug use of Tibetan medicine in treatment of gout disease (Zhihenai) based on data mining [J]. Clinical Research of Chinese Medicine, 2023, 15(35):120-124.