Analysis and Training of Network Information Document Management System Based on Data Mining

Wen Bo, Shi Li
School of Physics and Telecommunications Engineering
South China Normal University
Guangzhou, 510000
Guangdong, China
gbv2245374@163.com

ABSTRACT: *With the rapid development of information technology, document information management has become increasingly important. To improve the efficiency and accuracy of document information management, we propose a solution based on the BP neural network algorithm. This scheme first preprocesses the document information, including text cleaning, word segmentation, feature extraction, etc. Then, we used the BP neural network algorithm to classify and recognize the document information. Specifically, we used Multilayer Perceptron (MLP) as the model structure of the BP neural network algorithm, trained and optimized through a backpropagation algorithm. At the same time, we also used cross-validation and early stop techniques to avoid overfitting and underfitting issues. Through experimental verification, we found that the document information management system based on the BP neural network algorithm has high accuracy and efficiency. This system has higher classification accuracy and a lower false alarm rate than traditional text classification algorithms. In addition, the system also has good generalization performance and can adapt to the document information management needs of different fields.*

Keywords: Neural Network, Document Management, Information Management

Received: 3 March 2023, Revised 19 June 2023, Accepted 29 June 2023

DOI: 10.6025/jism/2023/13/3/70-77

1. Introduction

Information technology has now become a very important symbol of the development of various countries. Therefore, the circulation of information is getting more and more attention. How to manage information and reflect the variability and complexity of information circulation have become important elements in measuring the performance of an information system. Social information is multidimensional and complex, and population data information is more challenging to manage. Since data mining technology began in the 90s, its research has been extensive. The research scope involves association rule mining, classification rule mining, clustering rule mining, and trend analysis [1]. However, these studies are basically based

on structured data, such as the database of things, but few work to study heterogeneous and unstructured data. On the other hand, with the rapid development of the Internet, the network has developed into a distributed information space that has 300 million pages, which includes a large number of heterogeneous and unstructured information from technical data, commercial information to news reports and entertainment information and is still expanding [2]. Even industrial analysts believe that unstructured data accounts for 80% of the enterprise's information resources, while the data in the database accounts for only 20%. As a result, expanding the scope of data mining research and doing more research on unstructured data, such as text, web pages, Email and so on, have become a new research direction of data mining and network mining, text mining and multimedia mining emerge as the times require.

How to make full use of the document management information database has become a difficult problem in front of every clerical worker. The traditional way of data processing is simply manual statistics and query, summary and classification by computer. It is a data processing process for document transactions, which is closely related to the working experience of clerks and computer level. However, with the rapid growth of document data and the increasing number of document data, the original manual method has been unable to meet the needs of more complex documents in the new era. Therefore, it is becoming a new trend to discover the "document knowledge" hidden behind the data of the document by the computer.

2.State of the Art

We live in an era of Networked Information Technology, and communication, computer and network technology is changing human and the whole of society. A large amount of information brings convenience to people, but it also brings about the problem of excessive information and difficult to digest. With the rapid development of database technology and the wide application of database management systems, more and more data have been accumulated by people. A lot of important information is hidden behind the surging data, and people want to be able to analyze them at a higher level to make better use of these data. However, the current database system cannot find the relationships and rules in data and lacks the means of mining hidden knowledge behind data, which results in the phenomenon of "data explosion but poor knowledge".

There is an attribute-oriented reduction method that Canadian scholars propose. This method uses SQL-like language to express neural network queries, collect relevant data sets in the database, and then apply a series of data promotion technologies to data generalization on related data sets, including attribute deleting, concept tree lifting, attribute value control, counting and other aggregation functions and so on.

Not long after, other scholars put forward more perfect time series modelling theory and analysis methods. These classical mathematical methods predict time series by establishing stochastic models, such as the autoregressive model, autoregressive moving average model, summation autoregressive moving average model and seasonal adjustment model [3]. Because a large number of time series are not stationary, its characteristic parameters and data distribution vary with time.

Kohonen network is a typical self-organizing neural network, also called a self-organizing feature mapping network that is not to be denounced. Its input layer is a single neuron, while the output layer is a two-dimensional neuron; a lateral interaction exists between neurons as a "Mexico cap". Therefore, the output layer has a feedback characteristic between the neurons and the Kohonen network, which can be used as a pattern feature detector.

3.Methodology

3.1. BP Neural Network

A neural network is a new intelligent information processing theory developed in the process of imitating the problem of human brain processing. It consists of many simple processing units called neurons, constituting a nonlinear dynamical system. It stimulates and abstracts the brain's image thinking and associative memory to achieve information processing ability similar to the human brain's learning, recognition and memory. After more than 40 years of twists and turns, the neural network has shown great potential and broad application prospects in the field of information science and many other applications.

In the process of neural network development, the study of learning algorithms has a very important position [4]. Currently, the neural network models proposed by people all correspond to the learning algorithms. Therefore, there are sometimes no strict definitions or distinctions between models and algorithms. Some models can have a variety of algorithms, while others

may be used in various models.

In the neural network, the model samples provided by the external environment are trained, and the model can be stored, called the perceptron, which uses a teacher's signal to learn. The perceptron learning is the most typical learning of the neural network. A learning system with a teacher can be expressed in Figure 1; this kind of learning system can be divided into three parts: the input, the training and the output.

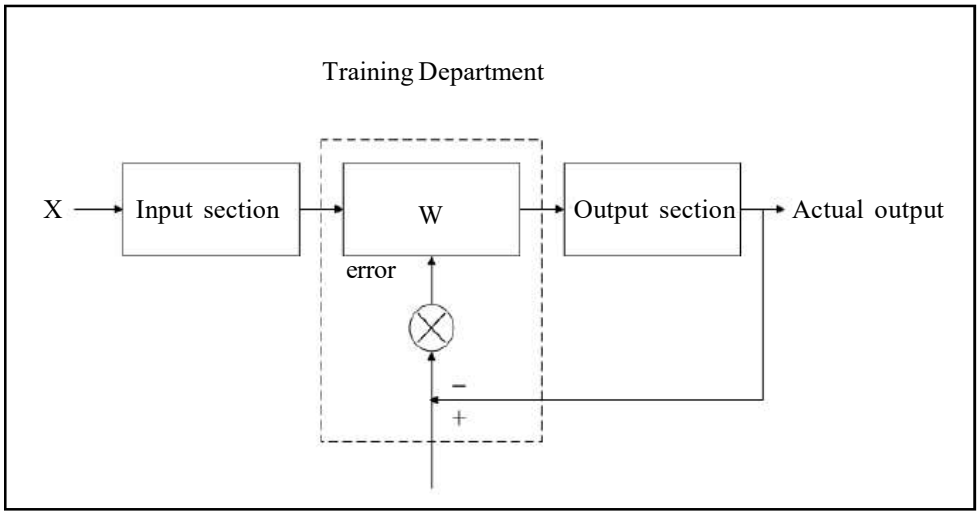

Figure 1. Block diagram of neural network learning system

The input unit receives a foreign input sample *X*, which is adjusted by the training department for the network weights and then outputs the results by the output. In this process, the desired output signal can be input as a teacher's signal, and the teacher's signal is compared with the actual output, resulting in the error of controlling the modified weight of the tooth.

Learning neural networks generally requires repeated training so that the error value is gradually approaching zero, so the learning of neural networks is consumed for a certain period. Some learning processes should be repeated many times, even tens of thousands of subordinates. Therefore, it is a very important research topic to improve the learning speed of the neural network and reduce the number of repetitions.

Text expression is mainly used as a vector space model(vSM). The basic idea of a vector space model is to represent a text with a vector: (W_1, W_2, W_3, K, W_0). Of which W_i is the weight of the *i* feature item, the feature item can generally choose words, words, or phrases. According to the experiment results, it is generally believed that the selection of words as a feature item is better than the word and phrase. Therefore, if we want to express text as a vector in vector space, we first need to segment text, and these words represent the text as the dimension of the vector, and the initial vector representation is 0 and 1 forms.

If this word appears in the text, then the dimension of the text vector is 1, otherwise 0. This method cannot reflect the degree of action of the word in the text, so 0 and 1 are gradually replaced by more accurate word frequency. Word frequency is divided into absolute word frequency and relative word frequency. The absolute frequency word, that is, the frequency of the use of words in the text to express text; and the relative word frequency is the normalized word frequency, and its calculation method mainly uses the formula TF-IDF. There are a variety of TF-IDF formulas, and we have adopted a more common TF-IDF formula in the system:

$$W_{ij} = \frac{tf_{ij} \text{ x } \log(N / n_i + 0.01)}{\sqrt{\sum_{K=1}^{N} [tf_{ij} \text{ x } \log(N / n_i + 0.01)]^2}} \qquad (1)$$

Among them W_{ij} is the weight of the word *I* in the text *J*, and tf_{ij} is the word frequency of the word *i* in the text *J*. *N* is the total number of training texts, and n_i is the number of texts in the training text, the denominator is the normalization factor.

The essence of the BP algorithm is to obtain the minimum value of the error function. This algorithm uses the fastest descent method in nonlinear programming and modifies the weight value according to the negative gradient direction of the error function. First, the error function *E* is defined, and the square sum of the difference between the expected output and the actual output is taken as an error function:

$$e = \frac{1}{2} \sum_i (X_i^m - Y_i)^2 \tag{2}$$

Among them, Y_i is the expected value of the output unit, which is used as a teacher signal; X^m is the actual output, because the m layer is the output layer. Since the BP algorithm modifies the weight value in the negative gradient direction of the error function e, the weight value W_{ij} is modified by ΔW_{ij} and e.

$$\Delta W_{ij} = - \eta \frac{\partial e}{\partial W_{ij}} \tag{3}$$

η is he learning rate, that is, the step length, and usually takes the number of [0-1]. Through repeated training of multiple samples, the weights are corrected in the direction of decreasing error to reach the final elimination error. It is also known from the above formula that if the number of layers of the network is large, the amount of calculation used is considerable, so the convergence speed is not fast.

The weight learning of neural networks is a complex continuous parameter optimization problem. If binary coding is adopted, then the coding string is too long and needs to be decoded as the real number to change the weight value into steps, affecting the network's learning accuracy [6]. Here, we use the real number code, as shown in Figure 2. Each weight value of the neural network is cascaded into a long string in a specific order, and each position on the string corresponds to a weight value of the network.

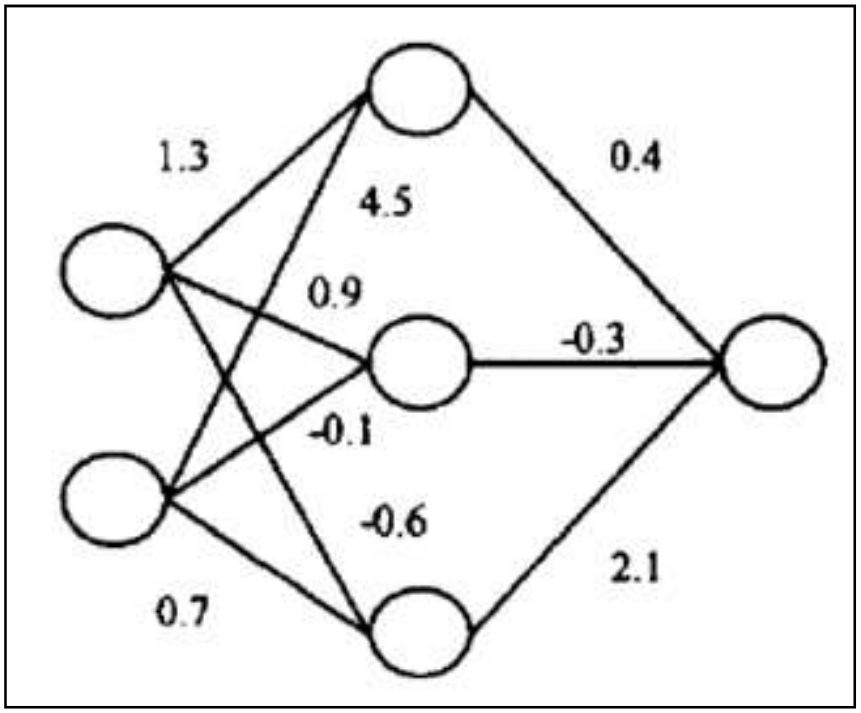

Figure 2. The coding method of weight learning problem in neural network

The initial concentration of each chromosome, the network weights are based on probability distribution to randomly determined, which is different with BP algorithm. In the BP algorithm, the initial weight is generally taken as a random number of uniform distribution between -1.0~1.0. The method of random distribution of genetic algorithm is obtained through a lot of experiments in the past. It can be found that the absolute value of weight is generally smaller after network convergence, but some of the weights are relatively large. The reason for using the above initialization method is to make the genetic algorithm be able to search the scope of all feasible solutions.

3.2. Design of Data Warehouse

The essential difference between data mining and traditional data analysis is that data mining is to excavate information and discover knowledge on the premise that there is no definite hypothesis. The information obtained from data mining should have three characteristics: pre-unknown, effective and practical. Data mining is driven by discovery, and the result is automatically extracted from the data through a large amount of analysis. That is to say, data mining is to discover information or knowledge that is not intuitively discovered or even to violate intuitive information or knowledge. The more unexpected information is, the more valuable it is [7].

The fact table is the core of the multidimensional model, which records business transactions and do index statistics. It is an information unit in a data warehouse, a unit in a multidimensional space, which is used to store data. According to different topics, different facts are designed as follows: document file archived directory volume fact table, which includes the main data of the archiving of each unit in the last four years, as shown in Table 1.

Field name	Field type	Field description
ID	Int	The primary key of the fact table
Gdyear_key	Int	Archival filing year
Ajtm_key	Int	Filing type of documents and Archives
Gddw_key	Int	Filing unit of documents and Archives
Ajscsj_key	Int	The generation of documents and Archives
Number of files	Int	The number of documents contained in the file volume

Table 1. Document File Archived Directory Number Fact Table

Developing a data warehouse system is a continuous growth and improvement process through continuous circulation and feedback. Its design mainly includes the model design and deployment and maintenance of the data warehouse. Star mode is the most commonly used data warehouse design structure implementation mode. It is composed of a fact table and a set of dimension tables, and each dimension table has a primary key, and all these dimensions constitute the primary key of the fact table [8]. The core of this pattern is the fact table, which connects various dimensional tables through the fact table, and each dimension table is connected to the central fact table. The following is the design of the dimension table with the document file data used as an example. The date dimension table is Dates, the corresponding table structure, as shown in Table 2, its dimension attributes constitute a conceptual layer.

On the basis of building a good file data warehouse, it is necessary to transfer the file data from the file information management system to the archival data warehouse. In addition, because of some historical and management reasons, there are many problems in the archives database. For example, file data description standard is not unified, man-made entry errors in data files, multiple files database inconsistencies in the field, the index exists null or duplicate values, code files are not unified, character format is not fixed, data format confusion, a large number of attributes is empty. These problems seriously affect the quality and effect of mining, and it is necessary to deal with these data before the data warehouse is built. Data separation is generally used to export data in transaction processing system to the temporary intermediate database by using SQL

statement according to certain standards and requirements, so as to carry out subsequent data processing, usually in the database. The following is an example of the archival filing catalogue data sheet of the archives of Tianjin University dbo.u_wswj, which mainly uses the SQL Server 2008 database for data access. The key to separating data is that it does not affect the regular operation of the file information management system as much as possible.

Field name	Field type	Field description
Date_key	int	The primary key of the dimension table, the external key of the fact table
Date	Datetime	Date of use of documents and Archives
Year	Char(4)	Year of use of documents and Archives
Quarter	int	Quarter of use of documents and Archives
Month	int	Month of use of documents and Archives

Table 2. Date Dimension Table

4.Result Analysis and Discussion

The first step in building a data warehouse is to determine its object and to build different types of data warehouses for other users. The theme is the key indicator of the data and interconnections involved in analysing the object. The division of the subject is mainly based on the analysis of the archives database and the interviews with the actual staff of the archives [9]. The existing archives database can well reflect the need for data analysis in the past file work, and the format and content of the file data are relatively stable and mature. In addition, we need to further explore potential user needs in daily work to have a wider and more comprehensive understanding of the theme division needed in constructing the archive data warehouse.

Data cleaning can be done by using SQL Server Integration Service (SSIS). SSIS is a platform for generating high performance data integration solutions, including the extraction, conversion, and load packages of data warehouses. The usual processing methods include merge, join, aggregate, sort, derive column, conditional split, row count, word search, word extraction, character mapping table and so on. SSIS data preprocessing is mainly carried out in the data flow module. The existing file directory data table has many empty values in multiple attributes (Table 3). Suppose a record contains a null value to delete this record. In that case, it may eventually lose information contained in a large number of actual data in the database and, at last, may get a smaller database, which changes the composition of the original database.

Content	**Archival year**	**Filing unit**	**File generation year**	**Paper title**
Null	198	109	130	36
Nonstandard	20	39	40	9

Table 3. Statistics of the Document Quality of the Tianjin University in the Last Four Years

When dealing with the problem of lack of value, it is usually more often filled with fixed value. For example, replace the classified level as the blank uniform with “inside” and the retention period with “long”. In the actual operation, different methods can be used to deal with the lack of value, and then the model is set up to compare each other separately, and the methods of high accuracy and low cost can be selected. The reason for the lack of value may be multifaceted. Some fields may be vacant, and filing Department staff input data files missing, but it may also be the file already does not have the contents of the field. For example, archives file numbers only have a specific number of superior talents, and ordinary schools have no fixed file transfer file number. This missing value indicates that the background of the document in the

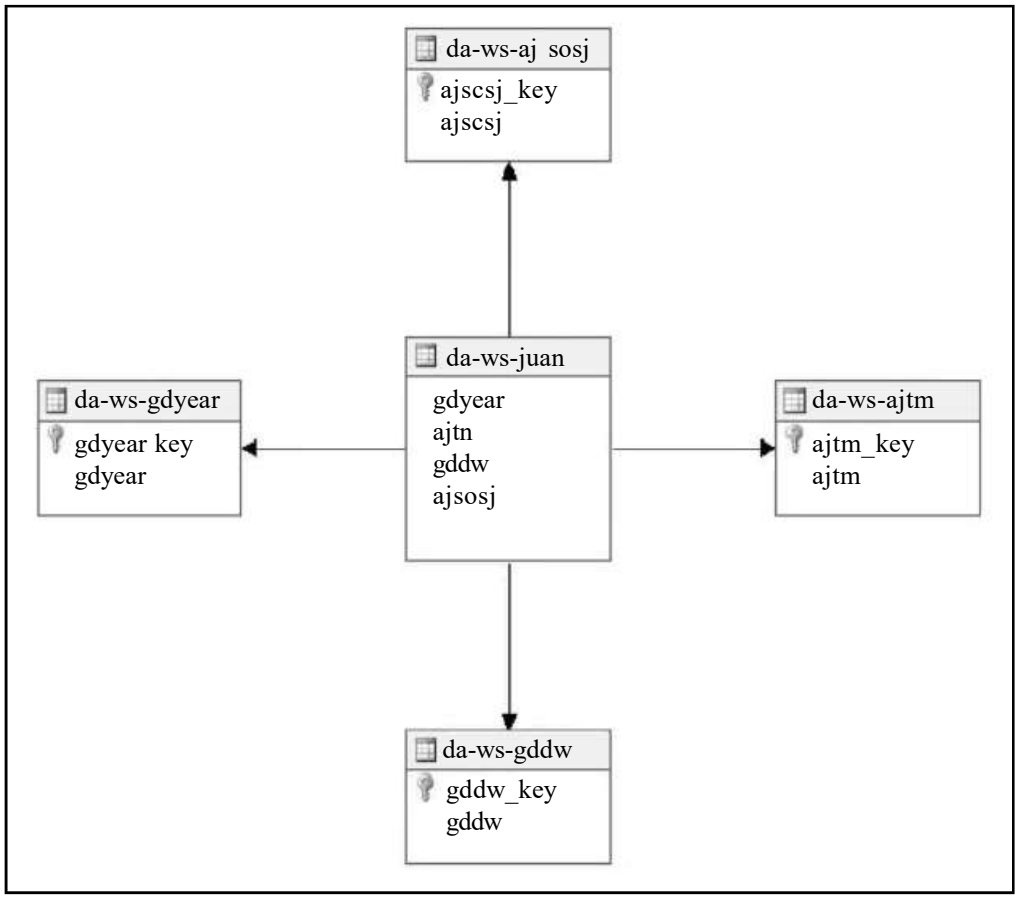

Figure 3. The coding method of weight learning problem in neural network

school is different from that of the superior document. We should treat the phenomenon of missing value of the archive data differently, especially when the data are formed [10]. For another example, if the file number, which sets the rules in the entry, such as the file number is empty when input "empty" when the file number properties appear to lack value, can be concluded by filing Department staff input errors.

5.Conclusion

Archival digitalization was the process of forming electronic documents by scanning or photographing the archives, which had made a certain improvement in the quantification of the archives. After analyzing and mining the data from the data tables of the documents and archives, it was found that many of the files with the word "superior text" in the title attribute were rarely used, which indicated that there was a problem in the whole process of archives collection in the initial identification of archives. The author drew the following conclusions. The cataloguing standard of documents and archives was not uniform and inaccurate. Many attribute data that were more suitable for data mining are all missing, especially the partial deletion caused by incomplete attribute settings in the register, which made the attribute with great excavation value useless and lost the significance of data mining. Most of the data collected in the front database needed to be manually recorded, all based on the entity files' various source data. However, a lot of data in archival practice management activities, such as web file data, file receiving data and so on, have not been collected into the system, or at present, there are not all kinds of electronic data by hand.

This article is in the primary stage of the study, and there are still some places to be perfected. For example, we should do as much as possible to collect a variety of data related to the work of pain files, and to improve the format and content requirements of various databases.

Acknowledgement

The authors would like to thank the editor and all the Anonymous reviewers for their valuable comments and suggestions that improved the paper's quality. This work is partially supported by the 2018 University Natural Science Foundation Project of Anhui Province, "Research on evaluation based on supply chain intelligent collaborative system", and the 2017 University Natural Science Foundation Project of Anhui Province, "Research on the Enterprise intelligent Efficiency evaluation under the Chinese intellectual background".

References

[1] Ding, S., Su, C., and Yu, J. (2011). An optimizing BP neural network algorithm based on genetic algorithm. *Artificial Intelligence Review*, 36 (2), 153-162.

[2] Zhang, L., Wu, K., Zhong, Y., et al. (2008). A new sub-pixel mapping algorithm based on a BP neural network with an observation model. *Neurocomputing*, 71 (10), 2046-2054.

[3] Yu, F., and Xu, X. (2014). A short-term load forecasting model of natural gas based on optimized genetic algorithm and improved BP neural network. *Applied Energy*, 134, 102-113.

[4]] Chau, K. W. (2007). Application of a PSO-based neural network in analysis of outcomes of construction claims. *Automation in construction*, 16 (5), 642-646.

[5] Ren, C., An, N., Wang, J., et al. (2014). Optimal parameters selection for BP neural network based on particle swarm optimization: A case study of wind speed forecasting. *Knowledge-Based Systems*, 2014, 56, 226-239.

[6] Shen, C., Wang, L., and Li, Q. (2007). Optimization of injection molding process parameters using combination of artificial neural network and genetic algorithm method. *Journal of Materials Processing Technology*, 183 (2), 412-418.

[7] Xia, C., Guo, C., and Shi, T. (2010). A neural-network-identifier and fuzzy-controller-based algorithm for dynamic decoupling control of permanent-magnet spherical motor. *IEEE Transactions on industrial electronics*, 57 (8), 2868-2878.

[8] Sedki, A., Ouazar, D., and El Mazoudi, E. (2009). Evolving neural network using real coded genetic algorithm for daily rainfall–runoff forecasting. *Expert Systems with Applications*, 36 (3), 4523-4527.

[9] Li, H. Z., Guo, S., Li, C. J., et al. (2013). A hybrid annual power load forecasting model based on generalized regression neural network with fruit fly optimization algorithm. *Knowledge-Based Systems*, 37, 378-387.

[10] Huang, J., Luo, H., Wang, H., et al. (2009). Prediction of time sequence based on GA-BP neural net. *Journal of University of Electronic Science and Technology of China*, 5, 029.