Application research of electronic archives information service system based on data mining technology

Bigeng Zheng1, *, Xinrui Chen2
1 Periodical Press of Jingchu University of Technology, Jingmen 448000, China
2 School of Electronics and Information Engineering, Jingchu University of Technology, Jingmen 448000, China
*Beginz@jcut.edu.cn

Abstract. With the continuous growth of colleges and universities, colleges and universities have accumulated a large number of precious archival resources. However, the current situation of archives management in colleges and universities is not optimistic, which not only has problems such as backward management, low utilization efficiency and insufficient information level. This paper analyzes the disadvantages of low utilization of information resources, irregular management and inadequate information service in archives management of colleges and universities. On the basis of exploring the information construction of archives in colleges and universities, the data mining technology is applied to the archives information service. Research and development of "university archives information service system based on data mining technology". This system is used to analyze the archival information of colleges and universities from multiple levels and angles, dig out the information rules and knowledge patterns hidden in the archival data, and constantly improve the standardization and intelligence level of archival management services.

1. Introduction
College archives are the original records formed in various work and activities such as enrollment, teaching, scientific research and academic research, which exist in different forms such as text and pictures [1-2]. It can intuitively reflect the process of colleges and universities at the beginning of their establishment and development to different periods. At the same time, it is also a reflection of the school's scientific research level, teachers and teaching experience. Archives in colleges and universities there exists backward management, use efficiency is low, such problems as insufficient information level, but also because of the lack of comprehensive information analysis and technical ability, also only stay in archives information management personnel for simple statistics, query, such as level, not on valuable precision classification and a deep analysis of archives information, a serious impediment to the modernization of the archives management of colleges and universities [3-4]. Archives repositories as an information resource center, is faced with serious challenges and the baptism of the information technology revolution, therefore, must improve the existing management mode, to raise efficiency of utilization of archives classification, and as well as the informationization level and so on aspects, the modern technology and the integration of archives management mode, and gradually turned to the service in order to Arthur as the center, change the original passive to active service service mode, to change the current situation of the development of archives in colleges and universities, is also the future development direction of college archives [5-6]. With the continuous increase of the country's informatization construction, and the release of the "ten-year development plan for educational
informatization (2011-2020)", the informatization construction is bound to play an important role in the education system, and will occupy a position that cannot be ignored, which is also an inevitable trend of the development of The Times[7-8]. But at present domestic universities archives system adopted by the general is still in the stage of entrust software company, is responsible for the development of technical personnel archives not generally professional background, the archives related knowledge and the actual operation process lack of thorough understanding of the software developed in such a case cannot meet the practical requirement of archives business completely. Or is to adopt the method of self-group research and development, but there are often problems such as backward technology. Therefore, it is urgent to have a complete, safe and efficient file service system to provide users with a full set of valuable file information[9-10].

With the rapid development and comprehensive utilization of computer technology, communication technology and multimedia technology, as well as the rapid development and popularization of science technology and informatization, human beings have entered the information age. But more and more data are found to be out of date before they can be analysed, or difficult to analyse because of the sheer volume of data. The lack of means to mine the hidden and undiscovered knowledge presents the embarrassing situation of "rich data but poor information". Therefore, how to analyze and process the obtained information resources in a more scientific, reasonable and planned way, and turn them into more beneficial resources for people? With the rapid development and popularization of computer technology, this ideal can become a reality.

The application of data mining technology can not only find useful and valuable knowledge and rules from a large amount of data, but also further analyze them, which is more convenient for people to understand and accept. Data mining technology is widely used in enterprises, business, insurance, medical care, manufacturing, engineering design and scientific exploration and other fields. The successful application in the commercial field brings great inspiration to university archives. If data mining technology is applied to archives work, it can not only improve the service utilization level and efficiency of archives, but also reduce the working intensity of archives managers. Especially in colleges and universities need to face different archives demand, and in the process of information accumulated a large number of data and other circumstances. How to provide better service, improve users' satisfaction, enhance the value of archives, and increase the social recognition of archives while meeting the demand of archivists is still a topic worthy of research.

The original idea of this paper is to change the status quo of archival information in colleges and universities and improve the utilization level and level of archival service. Give full play to the practical use and value of archives, stimulate the potential energy of archives, and maximize the efficiency of archives. Furthermore, it provides decision-making basis for the development and expansion of colleges and universities in a more efficient way, and plays an auxiliary role. At the same time, it also contributes to the development of archives in China.

2. Application of data mining technology in college archives information service
The existing archives system usually only completes the basic input of archives, information query, information change and other operations, but from the perspective of the development status of various colleges and universities, as well as the increasing amount of information, the original system cannot well analyze and utilize the useful data. Therefore, this system in the architecture, consider to apply data mining technology theory in the analysis of file system, through the analysis of the data mining technology in colleges and universities original file data, effective for archives information of scientific classification and summary, and then generate the corresponding rules and knowledge, to help workers college archives management to reduce costs, give full play to file its own resources, promote archives classification efficiency and service level, automation processing capabilities for archives information resources in colleges and universities, colleges and universities to achieve more efficient and accurate for leadership to provide decision-making data support, to promote the development of archive utility.
2.1. Application of data mining in archives

The daily work of university archives produces a large amount of data, and the complexity and diversity of the data are very high. This data is allowed to grow, and the end result is one "data grave" after another, which is like a chicken's ribs, occupying a large amount of computer resources, but must be kept and cannot be swept away. Without the intervention of a powerful data management and analysis tool, these data will lose its value for existence. Data mining technology is just such a tool, it can deal with massive data, mining the underlying laws behind the data, so that the potential value of data to fully reflect. Mining data generated by archives can provide technical support and decision management support for archives work.

2.2. Data mining process

Data mining is the process of discovering various models, profiles, and derived values from known data sets. Here, the word "process" is quite important. The data mining process is an iterative process, usually involving multiple interrelated steps, such as defining and analyzing topics, data preprocessing, selecting algorithms, extracting rules, evaluating and interpreting results, composing knowledge of patterns, and finally application. And the steps of data mining processing may vary depending on the application requirements and data base. In general, the basic steps of data mining include:

(1) Problem definition
(2) Establish a data mining model
(3) Analyze data
(4) Prepare data -- data select data conversion
(5) Model building
(6) Evaluation model
(7) Implement data mining data interpretation

To conduct data mining, first of all, the application field must be analyzed, including various knowledge and application objectives in the application. Understanding of the relevant field, familiar with the background knowledge, clear user requirements. Without background knowledge, it is impossible to clearly define the problem to be solved, to prepare good data for mining, and to interpret the results correctly. Clearly defining the business problem and recognizing the purpose of data mining is an important step in data mining. Understanding users' data and business problems is the first and most important thing before starting real data mining. Precisely defining the problem to be solved is one of the key factors for the success of data mining. In order to give full play to the value of data mining, it is necessary to have a clear and clear definition of users' goals, and an effective problem definition should also include a standard to measure the results of data mining. After determining users' needs, existing resources such as existing historical data should be evaluated to determine whether users' needs can be solved through data mining technology, and then the target of data mining and data mining plan will be further determined.

Data preparation: All the original data sets initially prepared for data mining are usually large, many of which are human-related and potentially messy. First, we should expect to find missing values, distortions, false records, and inappropriate samples in the initial data set. One of the most critical steps in data mining is the preparation and transformation of the initial data set. Raw data is not always the best data set for data mining, and many transformations have to be made to produce features that are more useful to the chosen data mining method. Applying different methods of calculation, using different sample sizes, selecting important ratios, and changing the size of the data window for time-dependent data, including changes in the moving average -- all these measures may contribute to better data mining results. Don't expect a machine to find the best set of transformations without human assistance, and don't expect a transformation used in one data-mining application to work well in another.

2.3. Decision tree algorithm

A particularly efficient way to generate a classifier from data is to generate a decision number. Decision tree representation is the most widely used logical method. Many decision tree induction algorithms
appear mainly in machine learning and applied statistical literature. They are guided learning methods that construct decision trees from a set of input-output samples. A typical decision tree learning system uses a top-down approach to search for solutions in a partial search space. It ensures a simple decision tree, but not necessarily the simplest. The decision tree includes nodes whose attributes have been validated. The output branching of a node corresponds to all possible test results for that node. Figure 1 shows a simple decision tree for sample classifications with two output attributes x and y. Samples of all attribute values x > 1, and y = B belong to class 2. A sample of value x < 1 belongs to class 1, regardless of the value of attribute y. For non-leaf nodes in the tree, you can continue to partition samples along branches. Each child node gets its corresponding subset of samples. The decision tree with single variable partition has a simple representation, and users can easily understand this reasoning model. At the same time, they show limitations on the expression of the model. Generally speaking. Any limitations on the expression of individual trees greatly limit the form of action and thus the model's ability to simulate. A well-known algorithm for generating decision trees is Quinlan's ID3 algorithm, which has an improved version called C4.5. Greedy search involves generating and pruning the structure of decision trees. It is typically used in power space algorithms to detect viable models.

![Decision Tree Diagram](image)

**Figure 1. A simple decision tree for testing attributes X and Y**

The ID3 algorithm starts with all the training samples at the root node of the tree. Select an attribute to partition these samples. A branch is generated for each value of the attribute, and the corresponding sample subset of the branch attribute value is moved to the newly generated child node. This algorithm is applied recursively to each child until all samples on a node are partitioned into a class. Each path to the leaves of the decision tree represents a classification rule. Note that the key decision of the top-down decision tree generation algorithm is the selection of node attribute values. The attribute selection of ID3 and C4.5 algorithms is based on minimizing the information entropy contained in the examples.

The method based on information theory insists on the minimum amount of data detected when classifying a sample in the database. The attribute selection of ID3 is based on an assumption that the complexity of the decision tree is closely related to the information expressed by the given attribute value. The information-based heuristic method selects the attribute that can give the highest information, that is, the attribute is the minimum information required for the result subtree of sample classification. The expansion algorithm of ID3 is C4.5 algorithm, which extends the classification range from classification attributes to digital attributes. The scale tends to use attributes that partition data into low-entropy subsets, where most samples belong to a single class. This algorithm basically selects the property that maximizes the local class differences.

### 2.4. C4.5 algorithm: generate a decision tree

The most important part of the C4.5 algorithm is the process of generating an initial decision tree from a set of training samples. Therefore, the algorithm generates a classifier in the form of a decision tree: a
structure with two types of nodes: a leaf node, representing a class, and a decision point, which specifies that the test is to be performed on a single attribute value, with a branch and subtree for each possible output of the test.

A decision tree can be used to classify a new sample, starting at the root of the tree and moving the sample to the leaves. At each non-leaf decision point, the attribute test results of the node are determined, and the attention is transferred to the root node of the selected subtree.

The architecture of C4.5 algorithm is based on hunter's CLS method, which constructs a decision tree through a group of training samples T. We use \{C1, C2, C3... Ck\} to represent these classes. The content information contained in set T has three possibilities:

T contains one or more samples. They all belong to a single class, Cj. Then the decision tree of T is a leaf node identified by class Cj.

T does not include samples. The decision tree is also a leaf, but the classes associated with that leaf are determined by information different from T, such as most classes in T. The C4.5 algorithm is guided by the classes that appear most frequently on the parents of the given node.

T contains samples of different classes. In this case, it's refining T to a subset of the sample set toward a single class of samples. Select output \{O1, O2,... On\}. T is partitioned into subsets T1,T2,... Tn, where the output of Ti including the test selected in T is all samples of Oi. The decision tree for T consists of a decision point identifying the validation and a branch for each possible output.

The same tree construction program is applied repeatedly to each subset of the training samples, so that Ti, the subset of the training samples, establishes the ith branch of the decision book. Continuous partitioning of training samples continues until all subsets contain samples belonging to the class.

The definition of the tree construction process is not unique. Different tests, even if they are applied in a different order, produce different trees. Ideally, we would like to check each step of the sample set partition to make the final tree smaller. Since we are looking for a compact decision tree consistent with the training set, why not find all the trees and choose the easiest one? Unfortunately, the problem of finding a minimum decision tree consistent with a training set is completely nonlinear. Enumerating and analyzing all possible trees leads to a combinatorial explosion of practical problems.

The original ID3 algorithm used a so-called gain criterion to select properties to test, based on the concept of entropy in information theory. The following relationship gives the weighted sum of the entropy of the set S (in bits):

\[
Info(S) = \sum_{i=1}^{k} \left( \frac{freq(C_i, S)}{|S|} \right) \log_2 \left( \frac{freq(C_i, S)}{|S|} \right)
\]

Now consider a similar metric after T is partitioned, and T partitions several outputs of X by examining one attribute. The required information can be obtained by the weighted sum of the entropy of these subsets:

\[
Info_x(T) = -\sum_{i=1}^{n} \left( \frac{|T_i|}{|T|} \right) \log_2 \left( \frac{|T_i|}{|T|} \right) * Info(T_i)
\]

The following quantities:

\[
Gain(X) = Info(T) - Info_x(T)
\]

The information obtained by measuring T partitioned according to test X is measured. This Gain criterion selects test X that maximizes Gain(X), that is, the criterion selects the attribute with the highest information Gain.

C4.5 algorithm generally contains three types of test structures:

The "standard" test for discrete values, with a branching and output for each possible value of the attribute.

If the attribute Y has consecutive values, define a binary check with the output Y<=Z and Y>Z by comparing this value with the threshold Z.

More complex tests based on discrete values in which each possible value of an attribute is assigned to a number of volatile groups, each with an output and branching.
While the gain criterion works well for the construction of a compact decision tree, it also has one serious drawback: it is severely biased against tests that have many outputs. The solution to this problem can be derived from several standardized methods. Specify an additional parameter, as defined by $\text{Info}(S)$ :

$$\text{Split} = \text{Info}(X) = -\sum_{i=1}^{n} (\frac{|T_i|}{|T|}) \log_2 \left(\frac{|T_i|}{|T|}\right)$$

This represents the potential information generated by partitioning the set $T$ into $n$ subsets $T_i$. Now define a new gain standard:

$$\text{Gain} = \text{ratio}(X) = \frac{\text{gain}(X)}{\text{Split} - \text{Info}(X)}$$

This new gain criterion represents the proportion of useful information generated by the partition, that is, it is helpful for classification. The gain ratio criterion also selects tests that maximize the ratio given earlier.

The same process should apply to other tests of decision trees. Maximising the gain rate replaces the gain criterion as the selection criterion for attribute selection and the test of subdividing samples. The final decision tree produced with the criteria of the new partition sample will be the most compact.

3. Analysis and design of archives information service system in colleges and universities

3.1. Feasibility analysis

(1) Technical aspects

With the rise of data mining technology, the rapid development and popularization of computer technology, communication technology and informatization, the research and development level of different kinds of operating systems has been rapidly promoted and developed. SQL Server series database launched by Microsoft has been liked and recognized by users in different industries and received good feedback. Therefore, the research and development of the system is feasible from the technical aspect.

(2) Economy

With the continuous development of computer technology, the cost of computer is getting lower and lower, and it is gradually accepted by all kinds of people in the society, almost reaching the state of popularization. The use of computer has been regarded as a basic skill in the work of various industries, and it is an indispensable part of office automation. The archives of colleges and universities have a large number of various and complex categories. For example, the artificial way is not only a waste of manpower, material resources and financial resources, but also a large amount of work and a long time. However, if computer technology is introduced into archive information management, it can save labor cost, time allocation and information management. Therefore, it is feasible to develop the system economically.

(3) Operation

College staff have a higher level of knowledge and education, in the computer operation is not a problem. The target system is developed according to the daily collection, classification, archiving, borrowing and other work processes of university archives. Each work process is represented by different modules. In this way, the system structure is clear and users with different identities only need to carry out relevant operations according to their own authority, which is simple and easy to understand.

From the above points, it can be concluded that the research and development of college archives information service system is feasible, and also is an effective way for college archives managers to improve the utilization rate of archives service.

3.2. System framework design

According to the demand analysis, this section discusses the overall architecture design of the content management subsystem of the image and electronic archive management system in detail.
The starting point for SSH framework planning is quite different from the existing modeling and component scheme design. The traditional website system design is: when the software designers and developers get the business requirements, they will immediately think whether to use Struts MVC structure, whether to use EJB or Hibernate, what kind of server and database to use, etc. Then the system architect communicates with each business person and divides each business module, and the business person fills in the corresponding code to each module. In this way, business is subordinate to technology, business functions are limited by specific technology, and business and technology are tightly coupled.

Such design makes many website systems limited by technology. Once the system needs to be improved or the technology is eliminated, their business will also change or be eliminated.

The SSH framework plan is designed to reduce the burden and effort of developers to rebuild solutions to complex problems; It can be extended for internal customization, improve development efficiency and easily achieve the scalability and maintainability of the system. The following is the hierarchy diagram of the system based on SSH framework:

![Figure 2. Application framework hierarchy diagram](image)

The hierarchical structure diagram shows the hierarchical relationship of application framework, which can be divided into persistence layer, logic layer, business layer, control layer and presentation layer. The upper component provides independent and complete functions for the lower component. The lower component does not need to know the implementation details of the upper component, but only needs to call the well-defined interfaces and methods provided by it to realize its own functions. The layered mechanism of application framework is responsible for dividing users' business requirements into relatively independent hierarchical modules, which enables developers to develop relatively independent business units easily and quickly.

Based on MVC pattern, the system is divided into presentation layer, middle layer (business logic layer) and data service layer. The three-tier system puts the work of business rules, data access and legitimacy verification in the middle layer. Instead of interacting directly with the database, the client connects to the middle tier through components, and the middle tier interacts with the database.

The presentation layer USES a jsp-based display with ExtJs component-based development mode.

The middle layer USES the popular Spring+Hibernate, and in order to separate the control layer from the business logic layer, it is further divided into the following categories.

The Web layer, which is the "C" (controller) in the MVC pattern, corresponds to the Action layer in the system. Responsible for controlling the interaction between the business logic layer and the
presentation layer, invoking the business logic layer, and returning business data to the presentation layer for organizational presentation.

The Service layer (the business logic layer) is responsible for implementing the business logic. Based on the DAO layer, the business logic layer completes the business logic required by the system through the front pattern packaging of DAO components.

The DAO layer, which is responsible for interacting with persistent objects. This layer encapsulates the operation of adding, deleting, checking and changing data, and adopts database middleware Hibernate to complete the encapsulation of the underlying database application. Through the consistent standard interface, the underlying database is separated from the business logic, providing data layer support for the business code development of the application system.

PO, persistent object. The data of relational database is mapped to objects by entity relational mapping tool, which makes it easy to operate database in an object-oriented way.

Spring functions throughout the entire middle layer, seamlessly integrating the Web layer, Service layer, DAO layer and PO. It relies on the dependency injection feature of Spring to inject Service in the Web layer, DAO in the Service layer, and Hibernate Template in the DAO layer, whose data Service layer is used to store data.

In order to improve the scalability and maintainability of the system, the interface layer and interface implementation layer are extracted in the DAO layer and Service layer respectively.

Figure 3 is the overall architecture of the electronic document management system.

\[
\text{Figure 3. The overall architecture of the content management subsystem of electronic document management}
\]

In terms of database middleware, the Dao layer was defined in the system, and the Hibernate Template integrated and packaged was used to implement data operations. The system USES the Annotation method to configure the mapping relationship between entity classes and database, including the primary key generation strategy, the configuration of the cascading relationship between tables and so on.

4. Conclusion

This article through to the current status of data mining technology at home and abroad are briefly introduced, on the basis of related concepts, algorithms, and through its application to the colleges and universities archives information service, technology to change the original work mode, put forward the feasibility of the implementation plan, application of archives information service system for mining and analysis of all kinds of archives in colleges and universities, digging out the improve the efficiency of archives classification, ascension archives management mode, and all kinds of factors, which restrict the development of the university archives and the development of archives in colleges and universities put forward its own ideas, in order to the current situation of the archives in colleges and universities improve and perfect.
Acknowledgements

2019 Jingmen Guiding Scientific Research Project “Research of Medical Image Encryption on Scrambling and composite segmentation chaotic mapping based on dynamic update”, Number: 2019YDKY078.

References

[1] Chu B H , Lee C E , Ho C S . An ontology-supported database refurbishing technique and its application in mining actionable troubleshooting rules from real-life databases [J]. Engineering Applications of Artificial Intelligence, 2008, 21(8):1430-1442.

[2] Wang J , Li L . Research on the College Graduate Employment Education Based on Data Mining Technology [J]. The Anthropologist, 2016, 23(1-2):231-235.

[3] Li Z , Hu Y Y , Zheng C Y , et al. Rules of meridians and acupoints selection in treatment of Parkinson's disease based on data mining techniques[J]. Chinese Journal of Integrative Medicine, 2018:1-5.

[4] Moreira M , Pelissari P I B G B , Parr C , et al. Data mining on technical trends and international collaborations in the refractory ceramic area[J]. Ceramics International, 2017, 43(9):6876-6884.

[5] Silva L F , Edoarda P S D R M , Andrade D A F R . ENADE: Math and Science Students\' Performance Analysis[J]. IEEE Latin America Transactions, 2017, 15(9):1742-1746.

[6] Peng X , Kong L , Chen D . A structural information-based twin-hypersphere support vector machine classifier[J]. International Journal of Machine Learning and Cybernetics, 2017, 8(1):295-308.

[7] Rong W , Guo S , Ding K , et al. Integrated strategy based on high-resolution mass spectrometry coupled with multiple data mining techniques for the metabolic profiling of Xanthoceras sorbifolia\'s Bunge husks in rat plasma, urine, and feces[J]. Journal of Separation Science, 2018.

[8] Kim E , Walter T , Powell J D . Wide Area Augmentation System-Based Flight Inspection System[J]. Journal of Aircraft, 2008, 45(2):614-621.

[9] Liu H , Yang C , Zhang M , et al. A computational intelligent approach to multi-factor analysis of violent crime information system[J]. Enterprise Information Systems, 2014, 11(2):161-184.

[10] Chen L , Liang X , Li T . Collaborative Performance Research on Multi-level Hospital Management Based on Synergy Entropy-HoQ[J]. Entropy, 2015, 17(4):2409-2431.