The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0264-0473.htm

Using data mining technology to solve classification problems

A case study of campus digital library

Chan-Chine Chang and Ruey-Shun Chen
Institute of Information Management, National Chiao Tung University, Hsinchu, Taiwan

Abstract

Purpose – Traditional library catalogs have become inefficient and inconvenient in assisting library users. Readers may spend a lot of time searching library materials via printed catalogs. Readers need an intelligent and innovative solution to overcome this problem. The paper seeks to examine data mining technology which is a good approach to fulfill readers' requirements.

Design/methodology/approach – Data mining is considered to be the non-trivial extraction of implicit, previously unknown, and potentially useful information from data. This paper analyzes readers' borrowing records using the techniques of data analysis, building a data warehouse, and data mining.

Findings – The paper finds that after mining data, readers can be classified into different groups according to the publications in which they are interested. Some people on the campus also have a greater preference for multimedia data.

Originality/value – The data mining results shows that all readers can be categorized into five clusters, and each cluster has its own characteristics. The frequency with which graduates and associate researchers borrow multimedia data is much higher. This phenomenon shows that these readers have a higher preference for accepting digitized publications. Also, the number of readers borrowing multimedia data has increased over the years. This trend indicates that readers preferences are gradually shifting towards reading digital publications.

Keywords Digital libraries, Electronic publishing, Knowledge mining, Multimedia

Paper type Case study

1. Introduction

The traditional library cannot satisfy customers with the same speed and convenience as a library with a computerized system (Yu and Chen, 2001). Therefore, it is essential that libraries have a smart and efficient way to help readers find useful books. Data mining is an important new information technology used to identify significant data from vast amounts of records. In other words, it is the process of exposing important hidden patterns in a set of data. It is also part of a process called knowledge discovery in databases, which presents and processes data to obtain knowledge. The usefulness of data mining is that it proactively seeks out trends within an industry and provides useful outcomes to organizations that maintain substantial amounts of information.

The goal of data mining is to improve the quality of the interaction between the library and its users. The collected data contain valuable information that can be integrated into the library's strategy, and can be used to improve library decisions. We need an automatic analysis and discovery tool for extracting useful knowledge from huge amounts of raw library data. Knowledge discovery in databases and the data

The Electronic Library
Vol. 24 No. 3, 2006
pp. 307-321

0264-0473
DOI 10.1108/02640470610671178

mining methodology are useful tools to apply to these objectives. Anand and Buchner (1997) claim that data mining can be defined as the process of extracting valid, previously unknown, comprehensible information from large databases in order to improve and optimize organizations' decisions. The term knowledge discovery in databases denotes the entire process of turning low-level data into high-level knowledge, where data mining is considered as a single step in the process that involves finding patterns in the data.

In this paper, we use data mining technology to elicit knowledge from databases and establish various kinds of data cubes, which will expand and aggregate data hierarchically to extract unknown information for decision-making purposes. Essentially, combinations of data mining and online analytical processing are used for data analysis to generate analytical results. Decision-makers then transform these results into graphs to develop important policies. This research used the following procedures using data mining and online analytical processing technology:

- confirmation of the goals of data mining – determine the problems to be solved by data mining;
- data selection – select the data from library massive databases;
- data processing – data cleaning, error removal, and data format consistency;
- data transformation – format adjustment, the joining or division of data fields;
- data storage – storing the data in an appropriate data repository;
- data dredging – classifying, sorting, and aggregating data to discover patterns and rules in order to assist decision-makers in making vital decisions; and
- user-relevant feedback – apply data mining mechanism and deliver query results to users: users then respond to the results.

2. Literature review

2.1 Digital library

Roger W. Christian was the first person to bring up an idea of the "electronic library" in 1975. F.W. Lancaster, in his book *Toward Paperless Information Systems* (Lancaster, 1978), predicted that electronic publication would replace paper publication after the year 2000, and that traditional libraries would shift to becoming digital libraries.

A digital library can provide a single point of access to a huge quantity of structured and accessible information that is available to a variety of users with different information needs. Digital libraries are inherently interactive systems with a constant growth of the number of end-users. They must not only rely on effective and sophisticated retrieval mechanisms, but must also provide efficient interaction with end users (Mulhem and Nigay, 1996).

Borgman (1999a, b) points out that digital libraries are a set of electronic resources and associated technical capabilities for creating, searching and using information. In this sense they are an extension and enhancement of information storage and retrieval systems that manipulate digital data in any medium and exist in distributed networks. The content of digital libraries includes data, metadata that describes various aspects of the data, and metadata that consists of links or relationships to other data or metadata, whether internal or external to the digital library.

Nowadays, the digital library has some new characteristics. Moyo (2004) points out a sample of typical user expectations in the electronic library environment:

- everything in full text and downloadable or printable;
- faster service;
- 24/7 service availability;
- easy access;
- virtual reference service librarian available online 24/7;
- easy-to-use web resources that permit self-service;
- a librarian who knows all subjects and all databases;
- everything in an electronic format;
- several options/alternatives to choose from;
- a website that works;
- the ability to conduct all library transactions online; and
- a website search engine that can find what the user wants.

Byrne (2003) states that the school digital library includes the following elements:

- *Integrated content provision*. The school digital library should support a wide range of digital resources, such as integration of the delivery of databases, e-journals and e-books, cross-file searching, and linking to full text and other services.
- *Support and training*. A school digital library should provide a digital environment to support and train users. It contains online real-time reference services, the provision of productivity software, and the creation of a website to promote the services.
- *Library effectiveness*. This include knowledge management support and the creation of websites to promote inter-library collaboration.

Byrne (2003) considers that this list of elements is not comprehensive, but shows the range of the characteristics of one university library.

2.2 Knowledge discovery in databases

The overall knowledge discovery in databases process is outlined in Figure 1. It is interactive and iterative, involving the following steps:

- *Step 1*. Developing an understanding of the application domain and the relevant prior knowledge. Identifying the goal of the knowledge discovery in databases process from the customer's viewpoint.
- *Step 2*. Creating a target data set: selecting a data set, or focusing on a subset of variables or data samples on which discovery is to be performed.

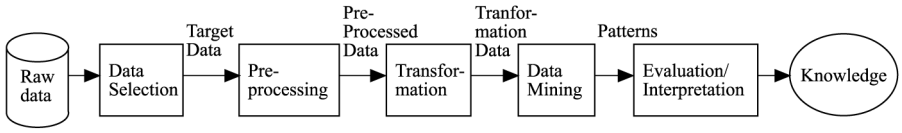

Figure 1.
Knowledge discovery in databases

- *Step 3*. Data cleaning and pre-processing: basic operations include removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes.
- *Step 4*. Data reduction and projection: finding useful features to represent the data depending on the goal of the task. With dimensionality reduction or transformation methods, the effective number of variables under consideration can be reduced, or invariant representations for the data can be found.
- *Step 5*. Matching the goals of the knowledge discovery in databases process (Step 1) to one of the particular data-mining methods, such as summarization, classification, regression, clustering, and so on. These data mining functions are described later in paper, and also by Fayyad *et al.* (1996).
- *Step 6*. Exploratory analysis, model and hypothesis selection: choosing the data mining algorithm(s) and selecting method(s) to be used for searching for data patterns. This process consists of deciding appropriate models and parameters and matching a particular data-mining method with the overall criteria of the knowledge discovery in databases process.
- *Step 7*. Data mining: searching for the desired patterns in a particular representational form or in a set of representations, such as classification rules or trees, regression, and clustering. A user can significantly apply the data-mining method by performing the preceding steps correctly.
- *Step 8*. Interpreting mined patterns, possibly returning to any previous step for further iterations. It is also possible to involve visualization of the extracted patterns and models or visualization of the data given the extracted models in this step.
- *Step 9*. Acting on the discovered knowledge: using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties. This process also includes checking and resolving potential conflicts with previously believed knowledge.

The knowledge discovery in databases process can involve significant iterations and can also contain loops between any two steps. Most previous work on knowledge discovery in databases has focused on Step 7. However, the other steps are just as important as Step 7 for the successful application of knowledge discovery in databases in practice.

2.3 Data mining

Knowledge discovery in databases refers to the overall process of turning low-level data into high-level knowledge. An important step in the knowledge discovery in databases process is data mining. Groth (2000) points out that data mining is the process of finding trends and patterns in data. The objective of this process is to sort large quantities of data and discover new information. The benefit of data mining is to turn this new-found knowledge into actionable results, such as increasing a customer's likelihood to buy, or decreasing the number of fraudulent claims. Berry and Linoff (1997, 1999) point out that data mining is the exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules. Yu and Chen (2001) note that the work process of data mining is composed of eight primary tasks (Figure 2).

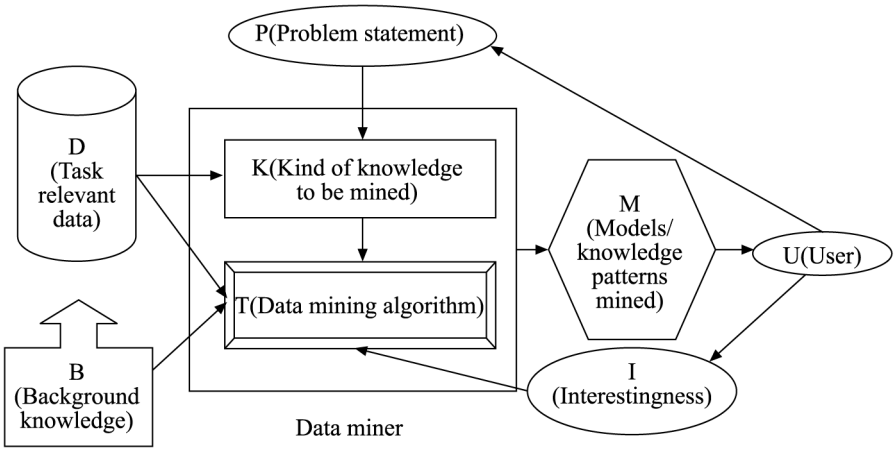

Figure 2.
Work process of data mining

The goal of data mining is to extract valuable and new information from existing data. Generally speaking, data mining includes the following major functions: classification, clustering, estimation, prediction, affinity grouping, description, etc. Data mining technology can be divided between traditional and refined technologies. Statistical analysis is representative of traditional technology. As for refined data mining technologies, all types of artificial intelligence are put to use. More commonly used types include decision trees, neural networks, genetic algorithms, fuzzy logic, rules induction, etc. The use of various types of application and different subjects of application can often lead to radically divergent results.

A particular data-mining algorithm is usually an instantiation of the model search preference components. After reviewing the articles by Noda *et al.* (1999), Han and Kamber (2001), and Buja and Lee (2001), we determined the most common model functions in current data mining practice to be the following:

- classification – classifies a data item into one of several predefined categorical classes;
- regression – maps a data item into a real valued prediction variable;
- clustering – maps a data item into one of several clusters, where clusters are natural groupings of data items based on similarity metrics or probability density models;
- rule generation – extracts classification rules from the data;
- discovering association rules – describes association relationships among different attributes;
- summarization – provides a compact description for a subset of data;
- dependency modeling – describes significant dependencies among variables; and
- sequence analysis – models sequential patterns, like time-series analysis.

Mitchell (1999) considers the impact of data mining to be due to:

- the falling cost of large storage devices and increasing ease of collecting data over networks;

- the development of robust and efficient machine learning algorithms to process this data; and
- the falling cost of computational power, enabling use of computationally intensive methods for data analysis.

3. Research methodology

This paper analyzes the borrowing records of readers in a campus library using the following techniques:

- data analysis;
- building a data warehouse; and
- data mining.

These techniques will be explained in more detail later in the paper.

3.1 Data collection and analysis

The following provides a brief description of the data set "Books loan registration data":

- *Data set description*. The data set contains reader data and reader's loan registration records between 2000 and 2003 in the digital library on campus. Reader's ID number, identification type, department and gender are covered in each reader's data. The loan registration record contains each reader's ID number, and time and detail information regarding book borrowing. The loan registration record has about million records. It is a very large data set; therefore, it should be analyzed and pre-processed with an efficient technique.
- *Data attribute analysis*. Definitions of the attributes of the tables "Reader" and "Borrow" are shown in Tables I and II.
- *Data pre-processing*. This describes the process of data collection for data mining, including data cleaning, data integration, data transformation and data reduction.

According to the mining process, the pre-processing is divided into two steps:

- *Step 1 – all data pre-process.* In this step, the pre-processing contains attribute removal, missing data, noisy data and inconsistent data, as shown in Tables III-VI.
- *Step 2 – focus on mining purpose.* In this step, we build the data relation and the data contrast on the data mining purpose, as shown in Figure 3.

Table I.
Attribute definition of "Reader" table

Item	Attribute name	Attribute	Type	Data value set
1	SEQNO	Sequential number	Continuous	Virtual sequential number
2	Dept_Code	Department	Nominal	000-YMT (385 types)
3	Type_Code	Identity type	Nominal	A-Q (16 types)
4	Sex	Sex	Nominal	Female, male

Item	Attribute name	Attribute	Type	Data value set
1	SEQNO	Sequential number	Continuous	Virtual sequential number
2	PROCESS_DATE	Process date	Date	Operation type: mm/dd/yy + time
3	ACC_NO	Register number	Continuous	000006-X285683
4	MARC_ID	Type number	Continuous	5-394,161
5	MARC_TYPE	Book type	Normal	0, 2
6	PUBLISH_YEAR	Publish year	Continuous	1909-2003, 9,999
7	SUBJECT	Subject	Normal	12,434 types
8	MARC_CLASS	Series category number	Continuous	Series category number
9	BOOK_CLASS	Book_ classification	Normal	000-990, A-Z
10	BOOK_CLASSTITLE	Classification_title	Normal	Text data

Table II.
Attribute definition of "Borrow" table

Item	Attribute name	Attribute	Type	Data value set
1	SNO	Key	Continuous	Sequential number
2	SEQNO	Sequential number	Continuous	Virtual sequential number
3	CODE	Book_classification	Nominal	100-900, A-Z
4	PROCESS_DATE	Process date	Date	Operation type: mm/dd/yy + time

Table III.
Attribute definition of "LibMain" table

Item	Attribute name	Attribute	Type	Data value set
1	SEQNO	Sequential number	Continuous	Virtual sequential number
2	Dept_Code	Department	Nominal	000-YMT (385 types)
3	Type_Code	Identity type	Nominal	A-Q (16 types)
4	Sex	Sex	Nominal	Female, male

Table IV.
Attribute definition of "Reader" table

Item	Attribute name	Attribute	Type	Data value set
1	SEQNO	Sequential number	Continuous	Virtual sequential number
2	SPECIAL_CODE	Record_Type	Normal	Operation type: mm/dd/yy + time
3	LAN	Language	Normal	eng/chi/...
4	PUBLISH_YEAR	Publish year	Continuous	1909-2003, 9,999
5	MARC_CLASS	Series category number	Continuous	Series category number
6	BOOK_CLASS	Book_ classification	Normal	000-900, A-Z
7	BOOK_CLASSTITLE	Classification_title	Normal	Text data

Table V.
Attribute definition of "BorrowDetail" table

Item	Attribute name	Attribute	Type	Data value set
1	CODE	Book_classification	Normal	000-900, A-Z
2	CLASS	Classification_title	Normal	Text data

Table VI.
Attribute definition of "Code" table

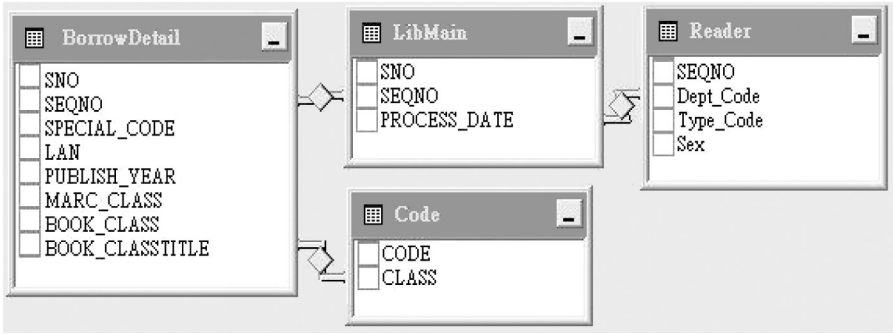

Figure 3.
The relations between data

3.2 Building a data warehouse

There are four steps in the process of building a data warehouse:

(1) *Setting up a schema for a data warehouse*. When establishing a data warehouse, the following frameworks can be utilized: star schema, snowflake schema and star flake schema. These three types of schema are all based on a fact table. The differences between them are their mutual relationships with external dimension tables. The dimension table in the star schema merely creates a connection with the library fact table, while different dimension tables have no relationship with each other. This research utilizes the star schema in designing the schema for the data warehouse. This schema is based on a Lib Fact table, a borrowing dimension table, a reader dimension table, and a classification dimension table. This is illustrated in Figure 4.

(2) *Setting up a Lib Fact table*. Real data are placed in the Lib Fact table. The data in this table cannot be altered; only new information can be added. Moreover, this table includes an index key related to other dimension tables. When designing a Lib Fact table, several factors must be taken into consideration:

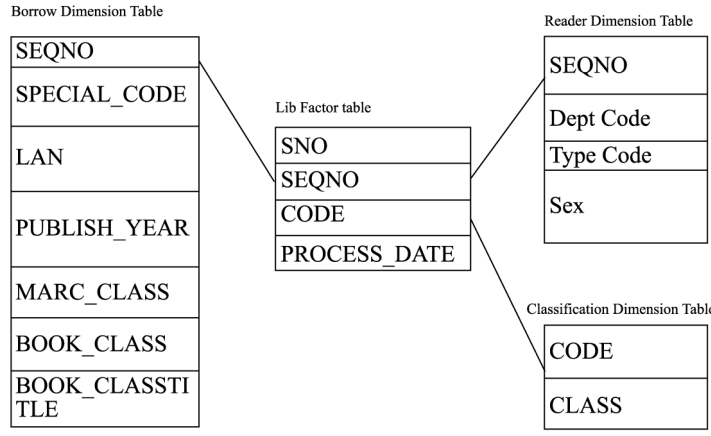

Figure 4.
Setting up the star schema for the data warehouse

- Determine which data are real and which data are dimensional.
- Decide a data warehouse period for all functions to achieve a balance between high-speed search capacity and data storage capacity. A long period of time is not necessarily positive. In fact, more precise data builds a better data warehouse. The time periods established for the data warehouse in this research include several years to measure trends in readers' data.
- Determine a principle to be used in a statistical sampling for all functions. Only a part of the real data should be placed in the data warehouse. Next, collected data are calculated according to the determined sampling principle.
- Determine which fields are included in the fact table and eliminate unused data; for example, status display fields, storage result fields and certain fields are used as internal references.
- To save space effectively for significant data, the size of the fields included in the fact table should be minimized.
- Determine whether or not to use an intelligent key to speed up the data search process.

(3) *Setting up a dimension table*. The dimension table data is used as a reference to the fact table data. If necessary, complex descriptions can be divided into several small parts, for example readers' information at a certain time. During the initial set-up stage, it is essential to assure that the dimension table's primary key will not be changed in any way. If the primary key changes, the fact table will also change. The dimension table is set up through a process of denormalization.

(4) *Setting up a multidimensional data model*. When analyzing data, multiple dimensions are brought together as one point of consideration. This process is called "multidimensional data modeling". Data warehouse systems may include many data cubes. Each data cube may be formed by different dimensions and fact tables. The online analytical processing operations in data cubes include roll-up, drill-down, slice, dice, and pivot (Chen, 2001). A data cube may be an *n*-dimensional data model. In order to provide an even wider range of search capabilities, we use the three dimensions – reader, code, and borrow – in this research to construct a three-dimensional data cube model as shown in Figure 5.

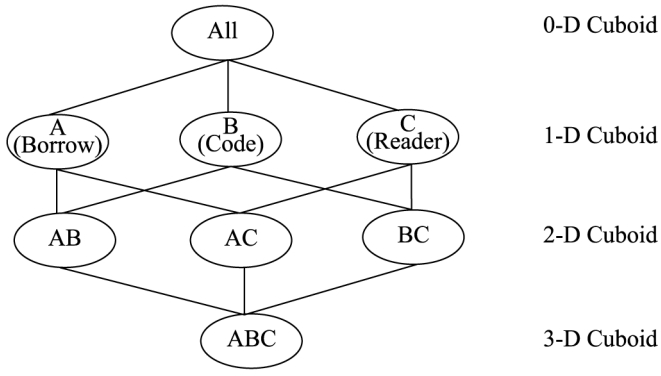

Figure 5.
Example of a three-dimensional data cube

3.3 Data mining process

The seven steps in the data mining process for library data are:

- *Step 1. Establish mining goals*. Deciding what the desired results are.
- *Step 2. Select data*. Deciding which data are useful, which attributes are worth considering, and how big the sample size should be.
- *Step 3. Pre-process data*. Filter out noisy, erroneous, or irrelevant data, and handle missing data.
- *Step 4. Transform data*. Where possible, reduce the number of data attributes or extract new ones from existing data attributes. Combine data tables and project the data onto working spaces – tables that represent the optimal abstraction level for the problem of interest.
- *Step 5. Store data*. Integrate and store data at a single site under a unified scheme.
- *Step 6. Mine data*. Perform appropriate data mining functions and algorithms according to mining goals. Typically, analysts first construct data cubes to provide multi-dimensional views of the data. Then they perform online analytical mining using the multi-dimensional data cube structure for knowledge discovery.
- *Step 7. Evaluate mining results*. Perform various operations such as knowledge filtering from the output, analyzing the usefulness of extracted knowledge, and presenting the results to the user for feedback. The feedback from this step can prompt changes to earlier steps.

4. Example: a case study for a campus digital library

This section details a practical case study for a campus digital library in which the data mining results form the data warehouse. The steps are described below:

- *Step 1. Establish mining goals*. In this research, we explore library readers' records and cluster readers by classifying their borrowing history.
- *Step 2. Select data*. Deciding which data is useful, which attributes are worth considering. The selected attributes are shown in Table VII.

Table VII. Selected attribute for mining

Attribute name	Data source table	Remark
SNO	LibMain	–
SEQNO	LibMain	Key
CODE	LibMain	–
SEQNO	Reader	Key
Type_Code	Reader	–
SEQNO	BorrowDetail	Key
SPECIAL_CODE	BorrowDetail	–
BOOK_CLASS	BorrowDetail	–
BOOK_CLASSTITLE	BorrowDetail	–
CODE	Code	Key
CLASS	Code	–

- *Step 3. Pre-process data*. Filter out noisy, erroneous, or irrelevant data, and handle missing data.
- *Step 4. Transform and store data*. In this step, we will build the data cube that utilizes the star schema in designing the schema for library data. This schema is based upon the Liball fact table, the reader dimension table, and the code dimension table. This is illustrated in Figure 6.

5. Mining data

Table VIII and Figure 7 show the mining result model. Five meaningful clusters can be found in the data set. In the data cube of books-borrowing records, the user classification is illustrated in Table IX and the borrowing content is categorized into several types shown in Table X. The Library of Congress Classification is used to classify foreign language books and the New Classification Scheme for Chinese Libraries is used to classify Chinese books. Each cluster includes types of readers, book classifications and other types of content (besides book).

6. Evaluation of mining results

We can distribute all readers into five clusters. There are different factors for each cluster, and these are described below:

(1) In Cluster 1, readers are undergraduates. They are interested in general works, and books about philology, linguistics, social science, natural science, history and Chinese geography, and philosophy.

(2) In Cluster 2, readers are graduates, associate researchers, and school staff. They are interested in books about military science, general legislative and executive

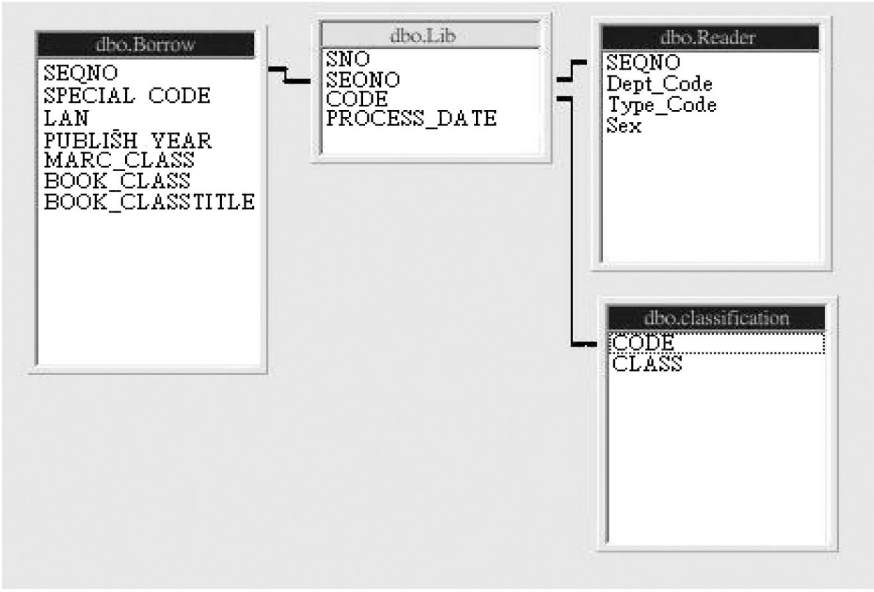

Figure 6.
Data cube

Cluster 1	Type code = A, Q, Book class = A, Book class = PA, Book class = PB, Book class = QS, Book class = 850, Book class = 080, Book class = 880, Book class = 820, Book class = 870, Book class = 830, Book class = 330, Book class = 780, Book class = 170, Book class = 600, Book class = 860, Book class = 590, Book class = 120, Special code = BOOK
Cluster 2	Type code = B, D, G, Book class = UA, Book class = JQ, Book class = HN, Book class = RE, Book class = PD, Book class = TT, Book class = HB, Book class = TX, Book class = 720, Book class = 660, Book class = 430, Book class = 550, Book class = 490, Book class = 480, Book class = 900, Book class = 920, Special code = BOOK
Cluster 3	Type code = A, B, Book class = AZ, Book class = AC, Book class = BD, Book class = BC, Book class = BJ, Book class = BP, Book class = DK, Book class = DU, Book class = F, Book class = GF, Book class = GC, Book class = HV, Book class = HJ, Book class = HX, Book class = JV, Book class = RM, Book class = RJ, Book class = TG, Book class = U, Book class = 650, Special code = BOOK
Cluster 4	Type code = A, N, H, Book class = HA, Book class = MT, Book class = NC, Book class = QA, Book class = PE, Book class = TK, Book class = T, Book class = 310, Book class = 440, Book class = 470, Book class = 510, Book class = 800, Book class = 960, Special code = CR, Special code = TA, Special code = HB, Special code = BOOK
Cluster 5	Type code = M, A, Book class = M, Book class = PN, Book class = PS, Book class = QU, Book class = 980, Book class = 910, Book class = 300, Book class = 990, Book class = 710, Book class = 970, Book class = 790, Book class = 520, Special code = LD, Special code = VCD, Special code = W, Special code = DVD, Special code = CD, Special code = VH, Special code = BOOK

Table VIII.
Data mining result

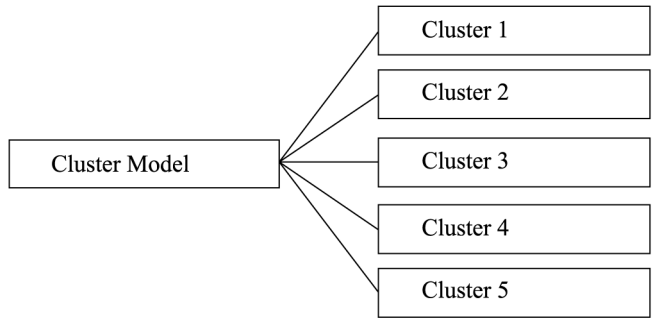

Figure 7.
The cluster model

papers, social science, medicine, philology, linguistics, technology, art and history, and geography of the world.

(3) In Cluster 3, readers are graduates, associate researchers, and undergraduates. They are interested in general works, books about philosophy, history, history of America, geography (atlases and maps), social science, general legislative and executive papers, medicine, technology, military science and history, and Chinese geography.

A	Undergraduate
B	Graduate/associate researcher
C	Teacher/researcher
D	Employee
E	NTHU/CCU/NYMU (graduate/teacher/employee)
F	Further education/volunteer/citizen
G	Inter-library loan/enterprise/exchange
H	Others
J	Credit course student
K	Practice assistant
L	Inter-library loan
M	Audio/video case
N	School fellow
P	Retired employee
Q	NTHU/CCU/NYMU (undergraduate)

Table IX.
User type code table

BOOK	Book
CD	CD
CR	CD (attachment with book)
CRM	CD (video)
DVD	DVD
HB	Hot book
LD	LD
R	Reference book
RB	Assign reference book
T	Thesis
TA	Tape
VCD	VCD
VH	VH
W	Writing of teacher

Table X.
Type code table

(4) In Cluster 4, readers are graduates, associate researchers, undergraduates, school fellows and others. They are interested in books about social sciences, music, visual arts, natural sciences, philology, linguistics, technology and art. Readers in this cluster also like to borrow CDs (attached to books) and tapes.

(5) In Cluster 5, readers are graduates and associate researchers. They are interested in books about music, philology, linguistics, science, natural science, social science, history, arts, and geography of the world. Readers in this cluster also like to borrow CDs, VCDs, LDs, DVDs, and VHS tapes.

Also, the statistical information shows that there are more people to borrow multimedia data from the library. The trend is shown in Figure 8.

7. Conclusion

Today, digital information is becoming ever more popular. The large quantity and diversity are the main features of digital information. Therefore, readers are interested in obtaining useful information efficiently. In this research, we aimed to achieve a significant outcome. We used data mining technology to discover some groups of

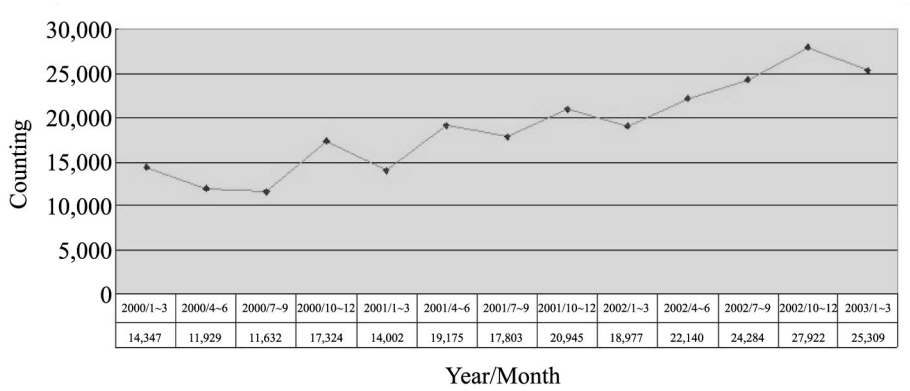

Figure 8.
Borrowing trend of multimedia data

readers from past borrowing records. The mining result shows that all readers can be categorized into five clusters, and each cluster has its own characteristics. Therefore, a digital library can anticipate a reader's needs in advance, depending on the mining results. We also discovered that the frequency of graduates and associate researchers borrowing multimedia data, such as CDs, VCDs, etc., is much higher. This phenomenon shows that these readers have a higher preference for accepting digitized publications. Also, according to the statistical information analyzed, we noticed that the number of readers borrowing multimedia has risen over the years. This upward trend indicates that readers are gradually shifting their reading preferences to digital publications.

References

Anand, S.S. and Buchner, A.G. (1997), *Decision Support Using Data Mining*, Prentice-Hall, Englewood Cliffs, NJ.

Berry, M.J.A. and Linoff, G.S. (1997), *Data Mining Techniques for Marketing, Sales, and Customer Support*, Wiley, New York, NY.

Berry, M.J.A. and Linoff, G.S. (1999), *Mastering Data Mining: The Art and Science of Customer Relationship Management*, Wiley, New York, NY.

Borgman, C.L. (1999a), "What are digital libraries? Competing visions", *Information Processing & Management*, Vol. 35, pp. 227-43.

Borgman, C.L. (1999b), "What are digital libraries, who is building them, and why?", in Aparac, T. (Ed.), *Digital Libraries: Interdisciplinary Concepts, Challenges and Opportunities*, Benja, Zagreb, p. 29.

Buja, A. and Lee, Y. (2001), "Data mining criteria for tree-based regression and classification", *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining San Francisco, CA*, pp. 27-36.

Byrne, A. (2003), "Digital libraries: barriers or gateways to scholarly information?", *The Electronic Library*, Vol. 21 No. 5, pp. 414-21.

Chen, Z. (2001), *Data Mining and Uncertain Reasoning: An Integrated Approach*, Wiley, New York, NY.

Fayyad, U., Shapiro, G.P. and Smyth, P. (1996), "The KDD process for extracting useful knowledge from volumes of data", *Communications of the ACM*, Vol. 39, pp. 27-34.

Groth, R. (2000), *Data Mining: Building Competitive Advantage*, Prentice-Hall, Englewood Cliffs, NJ.

Han, J. and Kamber, M. (2001), *Data Mining: Concepts And Techniques*, Morgan Kaufmann, San Mateo, CA.

Lancaster, F.W. (1978), *Towards Paperless Information Systems*, Academic Press, New York, NY.

Mitchell, T.M. (1999), "Machine learning and data mining", *Communications of the ACM*, Vol. 42 No. 11, pp. 30-6.

Moyo, L.M. (2004), "Electronic libraries and the emergence of new service paradigms", *The Electronic Library*, Vol. 22 No. 3, pp. 220-30.

Mulhem, P. and Nigay, L. (1996), "Interactive information retrieval systems: from user centered interface design to software design", *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, New York, NY, pp. 326-34.

Noda, E., Freitas, A.A. and Lopes, H.S. (1999), "Discovering interesting prediction rules with a genetic algorithm", *Proceedings of the IEEE Congress on Evolutionary Computing, CEC '99, Washington, DC, 6-9 July*, pp. 1322-9.

Yu, S.-C. and Chen, R.-S. (2001), "Developing an XML framework for an electronic document delivery system", *The Electronic Library*, Vol. 19 No. 2, pp. 102-10.

About the authors

Chan-Chine Chang is a graduate student in the PhD program in the Institute of Information Management of the National Chiao-Tung University in Taiwan.

Ruey-Shun Chen received his PhD in Computer Science and Engineering from National Chiao-Tung University, Taiwan in 1995. His research areas were internet applications and information systems. He is now an Associate Professor in the Institute of Information Management of National Chiao-Tung University in Taiwan. Ruey-Shun Chen is the corresponding author and can be contacted at: rschen@iim.nctu.edu.tw