DATA MINING AS A TECHNIQUE FOR BUSINESS PROCESS REDESIGN IN UNIVERSITY LIBRARY

Eka Miranda[1], Indrajani[2]
[1,2] *Department of Information System, Faculty of Computer, Binus University*
[1,2] *Jl. KH Syahdan no.9 Palmerah Jakarta Barat 11480*
[1] *ekamiranda@yahoo.com* , [2] *indrajani@yahoo.com*

ABSTRAK
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. The purpose of this paper is to suggest the use of data mining (DM) as a technique to support the process of redesigning a business by extracting the much-needed knowledge hidden in large volumes of data maintained by the organization through the DM models. Data mining is considered 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. The paper finds that after mining data, readers can be classified into different groups according to the publications in which they are interested. The data mining results shows that all readers can be categorized into three clusters; each cluster has its own characteristics. This phenomenon shows that these readers have a higher preference for accepting digitized publications.

Keywords*: Digital-libraries, Data-mining, Data-warehouse*

1. INTRODUCTION

The redesigning of an organization's processes is variously called business re-engineering, business process re-engineering, business process design, business redesign, and so on. A useful working definition of business process redesign (BPR) is the fundamental rethinking and radical redesign of an entire business – its processes, jobs, organizational structure, management systems, values and beliefs. BPR helps in rethinking a process in order to enhance its performance. Academics and business practitioners have been developing methodologies to support the application of BPR principles. However, most methodologies generally lack actual guidance on deriving a process design, thereby threatening the success of BPR. Indeed, a survey has proved that 85 per cent of projects fail or experience problems.

The traditional library cannot satisfy customers with the same speed and convenience as a library with a computerized system. 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 mining methodology are useful tools to apply to these objectives. 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:

a. confirmation of the goals of data mining – determine the problems to be solved by data mining
b. data selection – select the data from library massive databases

c. data processing – data cleaning, error removal, and data format consistency
d. data transformation – format adjustment, the joining or division of data fields
e. data storage – storing the data in an appropriate data repository
f. data dredging – classifying, sorting, and aggregating data to discover patterns and rules in order to assist decision-makers in making vital decisions
g. User-relevant feedback – apply data mining mechanism and deliver query results to users: users then respond to the results.

2. LITERATURE REVIEW

2.1 Business process redesign

When BPR is used carefully, it can take organizations into a new realm of competitive effectiveness. However, redesign of individual processes will always have a limited impact unless it is implemented as part of a wider view of the organization as a whole and that wider view must take root in the corporate culture. There is the difference between business re-engineering and process re-engineering since the first takes this wider perspective while the second is far more focused (Robson, 1997). The purpose of this paper is to present a DM technique that would allow business practitioners, senior managers and decision makers in organizations to extract useful, relevant, previously hidden knowledge from the organization's database, which after careful management, yields the knowledge needed to actualize the BPR.

2.2 The BPR framework

The idea behind a framework is to help practitioners by identifying the topics that should be considered and how these topics are related (Alter, 1999). In this perspective, the framework should identify clearly all views one should consider whenever applying a BPR implementation project.

For BPR, we suggest using the framework shown in Figure 1. In this framework, six elements are linked as shown in Figure 1.

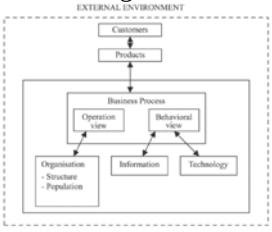

Figure 1 Framework for BPR implementation

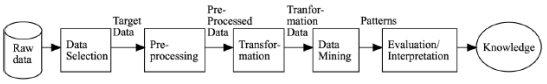

Figure 2 The overall knowledge discovery in databases process

2.3 Digital library

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.

Digital libraries are a set of electronic resources and associated technical capabilities for creating, searching and using information (Borgman, 1999a), (Borgman, 1999b). 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.

The university digital library includes the following elements (Byme, 2003) :

a. Integrated content provision. The university 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.
b. Support and training. The university 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.
c. Library effectiveness. This includes knowledge management support and the creation of websites to promote inter-library collaboration.

This list of elements is not comprehensive, but shows the range of the characteristics of one university library (Byme, 2003).

2.4 Knowledge discovery in databases

The overall knowledge discovery in databases process is outlined in Figure 2. It is interactive and iterative, involving the following steps :

a. 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.

b. 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 2. Knowledge discovery in Databases

c. 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.
d. Step 4. Data reduction and projection: finding useful features to represent the data depending on the goal of the task.
e. 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.
f. 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.
g. 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.
h. Step 8. Interpreting mined patterns, possibly returning to any previous step for further iterations.
i. 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.

2.5 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. Data mining is the process of finding trends and patterns in data (Groth, 2000). 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 newfound knowledge into actionable results, such as increasing a customer's likelihood to buy, or decreasing the number of fraudulent claims. 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 (Berry,Linoff, 1997), (Berry, Linoff, 1999). The work process of data mining is composed of eight primary tasks.

Figure 3. The work process of data mining

The goal of data mining is to extract valuable and new information from existing data. Data mining technology can be divided between traditional and refined technologies. Statistical analysis is representative of traditional technology.

A particular data-mining algorithm is usually an instantiation of the model search preference components. The most common model functions in current data mining practice to be the following:

a. classification – classifies a data item into one of several predefined categorical classes
b. regression – maps a data item into a real valued prediction variable
c. 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
d. rule generation – extracts classification rules from the data
e. discovering association rules – describes association relationships among different attributes
f. summarization – provides a compact description for a subset of data
g. dependency modeling – describes significant dependencies among variables
h. sequence analysis – models sequential patterns, like time-series analysis.

3. RESEARCH METHODOLOGY

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

a. data analysis
b. building a data warehouse
c. data mining

Figure 4. DM/BPR framework

3.1 Data collection and analysis

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

a. Data set description. The data set contains reader data and reader's loan registration records more than 10.000 in the digital library on university per year. memberID, fName, lName and detail information of each member are covered in each member's data. The loan registration record contains each transID, transDate, memberID, bookID, dateReturned, continueStatus 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.
b. Data attribute analysis. Definitions of the attributes of the tables "ms_Member" and "Trans" are shown in Tables 1 and 2.
c. 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 :

a. 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 3-10.
b. 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 5.

Table 1. Attribute definition of "ms_Member" table

Field Name	*Type*	*Size*	*Description*
memberID	Text	10	MemberID (Primary Key)
Password	Text	10	Password
fName	Text	20	First Name
lName	Text	25	Last Name
ktpID	Text	16	KTP (Identity Number)
address	Text	100	Address
cityID	Text	1	City ID
postCode	Text	5	Post Code
Phone	Text	8	Phone
Email	Text	30	Email Address
activeStatus	Boolean	-	If have already confirmed registration
activeDate	Date/Time	8	Active Date as a Member
paidStatus	Boolean	-	Paid Status

Tabel 2. Table 2. Attribute definition of "Trans" table

Field Name	*Type*	*Size*	*Description*
transID	Number	10	Transaction ID (Primary Key)
transDate	Date/Time	8	Transaction Date
memberID	Text	10	Member ID
bookID	Text	10	Book ID
dateReturned	Date/Time	8	Date Returned
continueStatus	Yes/No	-	Yes = have continue status, do not continue again

Table 3. Attribute definition of "ms_book" table

Field Name	*Type*	*Size*	*Description*
bookID	Text	10	Book ID (Primary key)
bookTitle	Text	100	Book Title
subject	Text	15	Subject of Book
author	Text	35	Author
publisher	Text	30	Publisher
review	Memo	-	Book's Overview
pict	Text	30	Hyperlink
tumb	Text	40	Hyperlink thumbnail

Table 4. Attribute definition of "trans_Book" table

Field Name	*Type*	*Size*	*Description*
bookID	Text	10	Book ID (Primary Key)
bookStatus	Text	10	OnLoan, Available, Booked
Onloan_till	Date/Time	8	Date Returned/Due Date

Table 5. Attribute definition of "City" table

Field Name	*Type*	*Size*	*Description*
cityID	Text	1	City ID (Primary Key)
cityName	Text	15	City Name

Table 6. Attribute definition of "temp_Trans" table

Field Name	*Type*	*Size*	*Description*
transID	Number	10	Transaction ID (Primary Key)
memberID	Text	10	Member ID
bookID	Text	10	Book ID

Table 7. Attribute definition of "Delivery" table

Field Name	*Type*	*Size*	*Description*
transDate	Date/Time	10	Transaction Date
memberID	Text	10	Member ID
bookID	Text	10	Book ID

Table 8. Attribute definition of "Take_In" table

Field Name	*Type*	*Size*	*Description*
transID	Number	10	Transaction ID
memberID	Text	10	Member ID
bookID	Text	10	Book ID
penalty	Number	8	Penalty

Table 9. Attribute definition of "Booking" table

Field Name	*Type*	*Size*	*Description*
transID	Number	10	Transaction ID (Primary Key)
memberID	Text	10	Member ID
bookID	Text	10	Book ID

Table 10. Attribute definition of "History" table

Field Name	*Type*	*Size*	*Description*
transID	Date/Time	10	Transaction Date
memberID	Text	10	Member ID
bookID	Text	10	Book ID

Table 11. Attribute definition of "ms_Category" table

Field Name	*Type*	*Size*	*Description*
categoryID	Text	10	Category ID (Primary Key)
categoryName	Text	50	Category Name

3.2 Building a data warehouse

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

a. Setting up a schema for a data warehouse. This research utilizes the star schema in designing the schema for the data warehouse. This schema is based on a Book Fact table, Member Fact table, City Fact table, Category Fact table, Take_in dimension table, and Trans dimension table. This is illustrated in Figure 5.

b. Setting up Book Fact table, Member Fact table, City Fact table, Category Fact table. Real data are placed in the Book Fact table, Member Fact table, City Fact table, Category 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 Book Fact table, Member Fact table, City Fact table, Category Fact table, several factors must be taken into consideration:

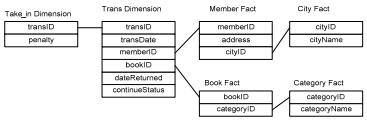

Figure 5. Setting up the star schema for the data warehouse

c. Determine which data are real and which data are dimensional.

d. Decide a data warehouse period for all functions to achieve a balance between high-speed search capacity and data storage capacity. The times established for the data warehouse in this research include several years to measure trends in readers' data.

e. 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.

f. Determine which fields are included in the fact table and eliminate unused data; for example, fName and lName fields, bookTitle fields and certain fields are used as internal references.

g. To save space effectively for significant data, the size of the fields included in the fact table should be minimized.

h. Determine whether to use an intelligent key to speed up the data search process.

i. 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 members' 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.

j. 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. A data cube may be an n-dimensional data model. In order to provide an even wider range of search capabilities, we use the two dimensions – Take_in and Trans – in this research to construct a two-dimensional data cube model as shown in Figure 6.

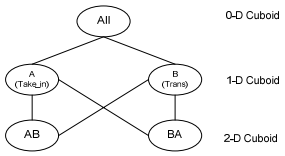

Figure 6. Example of a two-dimensional data cube

3.3 Data mining process

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

a. Step 1. Establish mining goals. Deciding what the desired results are.

b. Step 2. Select data. Deciding which data are useful, which attributes are worth considering, and how big the sample size should be.

c. Step 3. Pre-process data. Filter out noisy, erroneous, or irrelevant data, and handle missing data.

d. 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.

e. Step 5. Store data. Integrate and store data at a single site under a unified scheme.

f. 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.

g. 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 UNIVERSITY DIGITAL LIBRARY

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

a. Step 1. Establish mining goals. In this research, we explore library members' records and cluster members by classifying their borrowing history.
b. Step 2. Select data. Deciding which data is useful, which attributes are worth considering. The selected attributes are shown in Table 12.

Table 12. Selected attribute for mining

Attribute Name	*Data Source Table*	*Remark*
memberID	Member Fact	Key
address	Member Fact	
cityName	City Fact	
bookID	Book Fact	Key
categoryName	Category Fact	
transID	Trans Dimension	Key
memberID	Member Fact	Key
address	Member Fact	
cityName	City Fact	

c. Step 3. Pre-process data. Filter out noisy, erroneous, or irrelevant data, and handle missing data.
d. 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 Book Fact table, Member Fact table, City Fact table, Category Fact table, Take_in dimension table, and Trans dimension table. This is illustrated in Figure 7.

5. MINING DATA

Three meaningful clusters can be found in the data set. In the data cube of Trans records, the user classification is illustrated in Table 13 and the Trans content is categorized into several types shown in Table 14.

6. EVALUATION OF MINING RESULTS

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

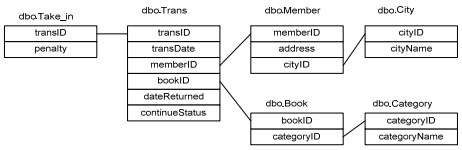

Figure 7. Data cube

In Cluster 1, readers are graduates and associate researchers. They are interested in general works, books about philosophy, history, history of the world, geography (atlases and maps), social science, general legislative and executive papers.

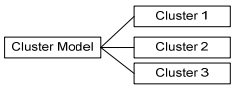

Figure 8. The cluster model

e. In Cluster 2, readers are graduates and associate researchers. 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.

f. In Cluster 3, 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.

Table 13. User Type Code

Code	*Description*
B	Graduate/associate researcher
C	Lecturer/researcher
D	Employee
J	Credit course student
K	Practice assistant
L	Inter-library loan
M	Audio/video case

Table 14. Type Code

Code	*Description*
BOOK	Book
CD	CD
CR	CD (sttachment 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 tearchers

7. CONCLUSION

As suggested in the DM/BPR framework, the DM model can be deployed on the massive data collected from past business processes of the organization, which then yields the previously unknown knowledge and trends needed by top managers or decision makers in the organization for effective BPR. The proposed DM/BPR framework transforms the old business into a new prospect oriented business organization by carefully re-engineering the old system incorporating the new discovered knowledge which helps the manager to make wise and informed business decisions in the area of accountability, business change management expertise, business process analysis, business model design, business model implementation and others.

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 readers from past borrowing records. The mining result shows that all readers can be categorized into three 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.

REFERENCES:

Alter, S. (1999), Information Systems: A Management Perspective, Addison-Wesley, Amsterdam.

Beriot, G. and Vemadat, F. (2001), "Enterprise modeling with CIMOSA: functional and organizational aspects", Production Planning and Control, Vol. 12 No. 2, pp. 128-36.

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.

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

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

Robson, W. (1997), Strategic Management and Information Systems: An Integrated Approach, 2nd ed., Financial Times Professional, London.