CloSpan Sequential Pattern Mining for Books Recommendation System in Petra Christian University Library

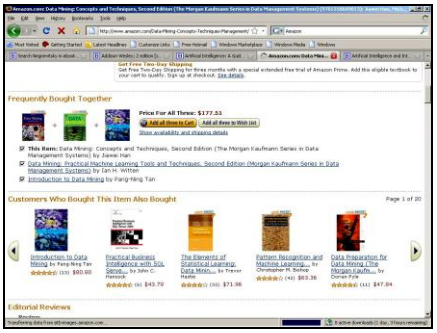
Gregorius S. Budhi[1], Yulia[2], Hery P. Gunawan[3]
Information Technology Department
Petra Christian University
Siwalankerto 121-131, Surabaya, East Java, Indonesia
greg@petra.ac.id, yulia@petra.ac.id

***Abstract*—Petra Christian University (PCU) Library has been using website for their books search system. To further improve the service, it is necessary to develop the automatic system which can recommends the book or the correlation or the book which often being lend at the same time or sequentially by prospective borrowers. The algorithm used to explore the lending sequential patterns is CloSpan Sequential Mining algorithm. The output generated by this application is closed sequential pattern rules and the tree of sequential patterns. They can be used as a reference to establish a list of recommended related books. From the test results it can be concluded that the more data and smaller minimum support, the longer the process takes, and the more patterns that is produced. From the questionnaire outcome that are distributed to employees and users of the library can be concluded that the system can create right recommendations and useful.**

***Keywords*—Data Mining, Sequential Pattern Mining, Clospan, Books Recommendations, Library**

I. Introduction

Until today Petra Christian University (PCU) library has been using website as a service to find books from their collections. Later this library intends to improve the service features provided by the website. One of it is to add recommendation features about other books that relate to the title of the book that is being viewed on the website. These relations are the books that are often borrowed together and the book that are often borrowing sequentially with the book that is viewed. This is similar to the service provided by the online bookstores such as Amazon.com (Picture 1).


Amazon.com: Data Mining: Concepts and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems)
File Edit View History Bookmarks Tools Help
http://www.amazon.com/Data-Mining-Concepts-Techniques-Management/
Amazon
Most Visited
Getting Started
Latest Headlines
Customize Links
Free Hotmail
Windows Marketplace
Windows Media
Windows
Amazon.com: Data Mini...
Add all three to Cart
Add all three to Wish List
Editorial Reviews
Transferring data from g-ecx.images-amazon.com...

Get Free Two-Day Shipping
Get Free Two-Day Shipping for three months with a special extended free trial of Amazon Prime. Add this eligible textbook to your cart to qualify. Sign up at checkout. See details.

Frequently Bought Together

Price For All Three: $177.51

Show availability and shipping details

☑ This Item: Data Mining: Concepts and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems) by Jiawei Han
☑ Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems) by Ian H. Witten
☑ Introduction to Data Mining by Pang-Ning Tan

Customers Who Bought This Item Also Bought — Page 1 of 20

Introduction to Data Mining by Pang-Ning Tan (33) $80.80
Practical Business Intelligence with SQL Serve... by John C. Hancock (6) $43.79
The Elements of Statistical Learning: Data Minin... by Trevor Hastie (30) $71.96
Pattern Recognition and Machine Learning... by Christopher M. Bishop (42) $63.36
Data Preparation for Data Mining (The Morgan Kaufm... by Dorian Pyle (11) $47.84

Figure 1. Amazon webpage

In order to build this recommendation service feature, researchers approached it with data mining method, namely sequential pattern mining. First it will be mined patterns of frequently borrowed books simultaneously and sequentially. The patterns that were found will be the basis of the recommendations. The algorithm used to dig patterns is CloSpan Sequential Pattern Mining. The output of the system can be viewed in the form of sequential pattern rules and a sequential patterns tree. To further improve the results of mining pattern, we dig data in a multidimensional manner. So that in addition to the book loan data, we entered the books borrower data such as, address, age, home department and the entry year (if a student).

This application is the development and improvement of a similar application that was created earlier using different sequential mining algorithms [2]. The CloSpan algorithm used in this application because theoretically this algorithm produces less number of rules but the same function and also has a faster processing time.

II. Basic Concept

A. *Sequential Pattern Mining*

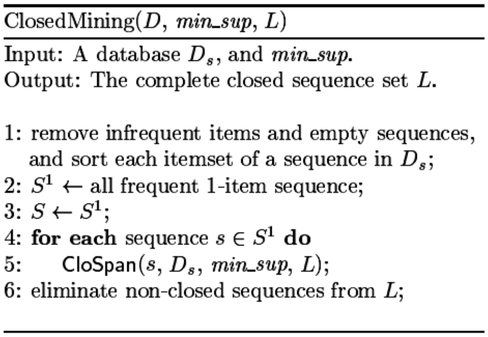
Sequential data is pervasive in our lives. For example, your schedule for a particular day is the sequence of your activity. When you read a story, you informed the development of some of the events that are also sequential. If you have an investment in the company, you are interested to learn the history of the company stocks. Deep in your life, you rely on biological sequences, including DNA and RNA sequences [1].

Given a set of sequences, where each sequence consists of a list of events (or elements) and each event consists of a set of items. Then given limit a user-specified minimum support, mining sequential patterns will found all frequent subsequences [4].

B. *Closed Sequential Patterns Algorithm (CloSpan)*

CloSpan is a method of sequential pattern mining. CloSpan can make the process of mining for long sequences. CloSpan produce significant sequences with smaller amounts rather than using the traditional method, which retains the same power of the overall expressive frequent subsequences [5].

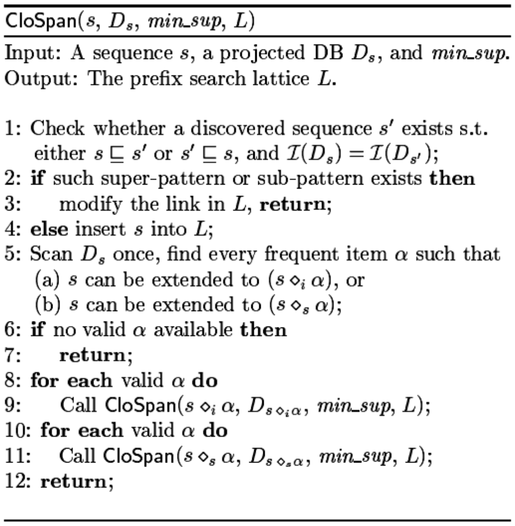
Broadly speaking, CloSpan consists of two main phases, namely: First phase, bring CloSpan LS set, a superset of the frequent closed sequences, and store it in a prefix sequence lattice, and second phase is post - pruning CloSpan to eliminate non-closed sequences. CloSpan algorithm can be seen in Figure 2 and Figure 3.



```
ClosedMining($D$, $min\_sup$, $L$)
Input: A database $D_s$, and $min\_sup$.
Output: The complete closed sequence set $L$.

remove infrequent items and empty sequences,
  and sort each itemset of a sequence in $D_s$;
$S^1 \leftarrow$ all frequent 1-item sequence;
$S \leftarrow S^1$;
for each sequence $s \in S^1$ do
    CloSpan($s$, $D_s$, $min\_sup$, $L$);
eliminate non-closed sequences from $L$;
```

Figure 2. Algorithm Closed Mining [5]



```
CloSpan($s$, $D_s$, $min\_sup$, $L$)
Input: A sequence $s$, a projected DB $D_s$, and $min\_sup$.
Output: The prefix search lattice $L$.

Check whether a discovered sequence $s'$ exists s.t.
  either $s \sqsubseteq s'$ or $s' \sqsubseteq s$, and $\mathcal{I}(D_s) = \mathcal{I}(D_{s'})$;
if such super-pattern or sub-pattern exists then
    modify the link in $L$, return;
else insert $s$ into $L$;
Scan $D_s$ once, find every frequent item $\alpha$ such that
  (a) $s$ can be extended to $(s \diamond_i \alpha)$, or
  (b) $s$ can be extended to $(s \diamond_s \alpha)$;
if no valid $\alpha$ available then
    return;
for each valid $\alpha$ do
    Call CloSpan($s \diamond_i \alpha$, $D_{s \diamond_i \alpha}$, $min\_sup$, $L$);
for each valid $\alpha$ do
    Call CloSpan($s \diamond_s \alpha$, $D_{s \diamond_s \alpha}$, $min\_sup$, $L$);
return;
```

Figure 3. CloSpan Algorithm [5]

C. Multidimensional Patterns Mining

This method allows extracting information from multiple attributes or dimensions, so that data mining is not only based on one attribute only. Mining process would be more effective if the main attribute can be associated with other attributes or dimensions, so the result can be seen from many aspects of the associated data [3].

Multidimensional process of pattern mining is as same as single dimensional pattern mining. The different is it was in the beginning it done merging the data with some desired dimensions. After that the dimensions can be considered as separate items, so it can be calculated as a single-dimensional pattern mining. The results obtained are in the form of multidimensional rules.

III. APPLICATION DESIGN

This application is further development of similar application has been made by researcher for PCU library [2]. The developments are as follows:

- The addition of multidimensional elements of the application. With the addition of a sequential pattern will be obtained not only single dimension (book-to-book), but also can be seen from some of the other aspects, e.g.: "If the borrower from the Informatics department will borrow the book Data Mining Concept and Technique next borrowed the book Pattern Discovery Using Sequence Data Mining."

- Changes in the algorithms used, namely: from algorithm PrefixSpan to CloSpan. The reason is CloSpan algorithm produces smaller number of sequential patterns but equally meaningful. These algorithms form a pattern that has only "close" or not to be sub-pattern from another pattern longer. Fewer numbers of patterns will further accelerate the search process when applied to the PCU library books search system.

The Block diagram of design of the application can be seen in Figure 4.

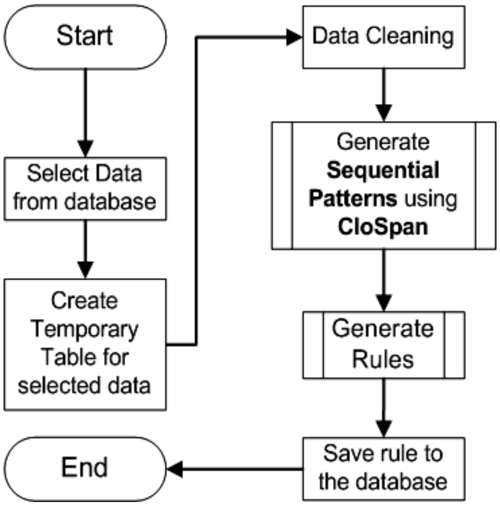
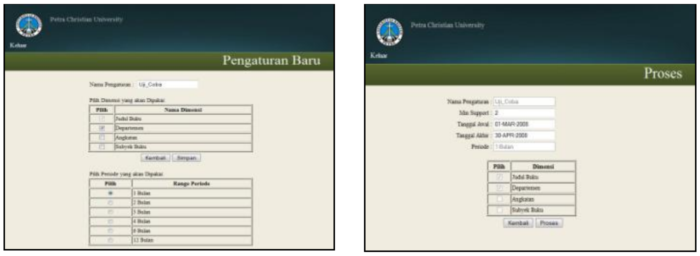

Start
Select Data from database
Create Temporary Table for selected data
Data Cleaning
Generate **Sequential Patterns** using **CloSpan**
Generate Rules
Save rule to the database
End

Figure 4. Block Diagram of the Application

Application starts with selecting the data to be mined, namely: Selecting dimension / attribute data to be processed, namely: book titles, department, student entry year and books subject. After that we can select data period between 1-12 months from the current month. Then the user can specify how much the minimum support is permitted. The webpage to select dimensions, period of the data and minimum support can be seen in Figure 5.



Petra Christian University

Keluar

Pengaturan Baru

Nama Pengaturan : Uji_Coba

Pilih Dimensi yang akan Dipakai

Pilih	Nama Dimensi
☐	Judul Buku
☑	Departemen
☐	Angkatan
☐	Subyek Buku

Pilih Periode yang akan Dipakai

Pilih	Range Periode
●	1 Bulan
○	2 Bulan
○	3 Bulan
○	4 Bulan
○	6 Bulan
○	12 Bulan

Petra Christian University

Keluar

Proses

Nama Pengaturan : Uji_Coba

Min Support : 2

Tanggal Awal : 01-MAR-2008

Tanggal Akhir : 30-APR-2008

Periode : 1 Bulan

Pilih	Dimensi
☑	Judul Buku
☑	Departemen
☐	Angkatan
☐	Subyek Buku

Figure 5. Pages to select dimensions, the period of data and specify minimum support.

Next, the application will create a temporary table to store the data that has been selected. The data is then going through a cleaning process where data that is lost or corrupted can be repaired. Cleaning process is divided into two parts, namely:

- Cleaning the data manually, which the administrator can repair or fill corrupted or missing data records.
- Cleaning the data automatically. For automated data cleaning there are some rules to follow, namely:
    - Book Title data, when data is empty then it will be filled with the value 0 (Unknown).
    - Data Department, if the field is empty then the application will automatically track member who borrowed it from his member code. This is because the digits 1 to 3 of the code are the code of members department.
    - Student entry year data can be traced also from the member code digit 4 and 5.

After the cleaning process complete the data is processed using CloSpan algorithm to forming frequent closed sequences. Flowchart of the CloSpan can be seen in Figure 6 and Figure 7.

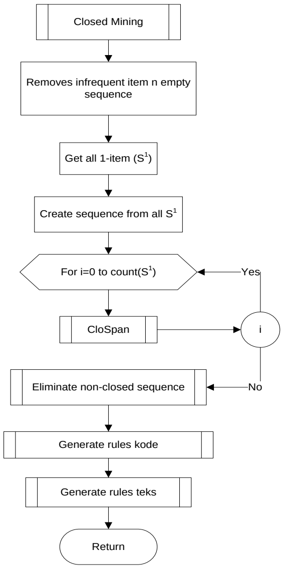


Figure 6. Closed Mining Process

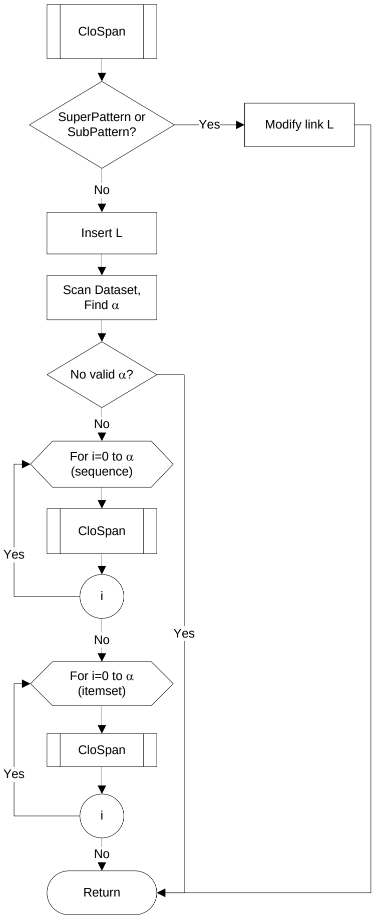
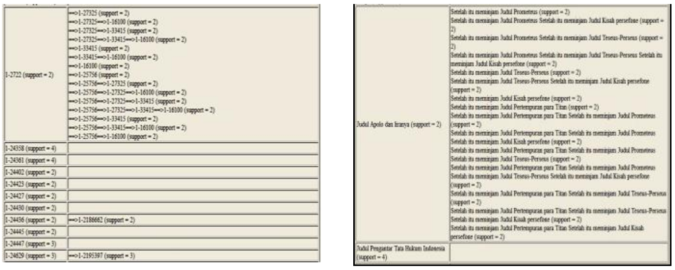


Figure 7. CloSpan Process

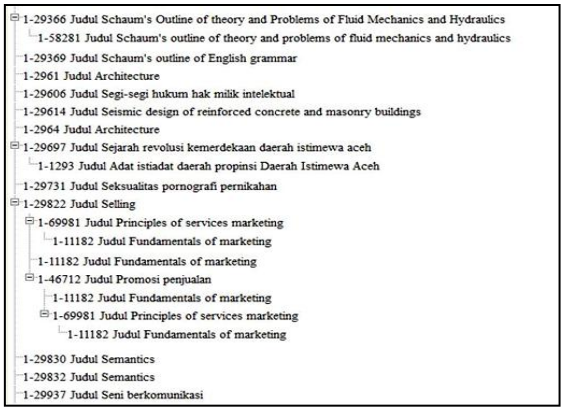
Then we built rules and tree from the closed frequent sequences. The rules and tree are for the visualization of the results of closed frequent sequences to the administrator or other users in need. Webpage of rules and tree visualization of frequent closed sequence is shown in Figure 8 and Figure 9.



1-2722 (support = 2)	==>1-27325 (support = 2) ==>1-27325==>1-16100 (support = 2) ==>1-27325==>1-33415 (support = 2) ==>1-27325==>1-33415==>1-16100 (support = 2) ==>1-33415 (support = 2) ==>1-33415==>1-16100 (support = 2) ==>1-16100 (support = 2) ==>1-25756 (support = 2) ==>1-25756==>1-27325 (support = 2) ==>1-25756==>1-27325==>1-16100 (support = 2) ==>1-25756==>1-27325==>1-33415 (support = 2) ==>1-25756==>1-27325==>1-33415==>1-16100 (support = 2) ==>1-25756==>1-33415 (support = 2) ==>1-25756==>1-33415==>1-16100 (support = 2) ==>1-25756==>1-16100 (support = 2)
1-24358 (support = 4)	
1-24361 (support = 4)	
1-24402 (support = 2)	
1-24423 (support = 2)	
1-24427 (support = 2)	
1-24430 (support = 2)	
1-24436 (support = 2)	==>1-2186662 (support = 2)
1-24445 (support = 2)	
1-24447 (support = 3)	
1-24629 (support = 3)	==>1-2195397 (support = 3)

Judul Apolo dan Iranya (support = 2)	Setelah itu meminjam Judul Prometeus (support = 2) Setelah itu meminjam Judul Prometeus Setelah itu meminjam Judul Kisah persefone (support = 2) Setelah itu meminjam Judul Prometeus Setelah itu meminjam Judul Teseus-Perseus (support = 2) Setelah itu meminjam Judul Prometeus Setelah itu meminjam Judul Teseus-Perseus Setelah itu meminjam Judul Kisah persefone (support = 2) Setelah itu meminjam Judul Teseus-Perseus (support = 2) Setelah itu meminjam Judul Teseus-Perseus Setelah itu meminjam Judul Kisah persefone (support = 2) Setelah itu meminjam Judul Kisah persefone (support = 2) Setelah itu meminjam Judul Pertempuran para Titan (support = 2) Setelah itu meminjam Judul Pertempuran para Titan Setelah itu meminjam Judul Prometeus (support = 2) Setelah itu meminjam Judul Pertempuran para Titan Setelah itu meminjam Judul Prometeus Setelah itu meminjam Judul Kisah persefone (support = 2) Setelah itu meminjam Judul Pertempuran para Titan Setelah itu meminjam Judul Prometeus Setelah itu meminjam Judul Teseus-Perseus (support = 2) Setelah itu meminjam Judul Pertempuran para Titan Setelah itu meminjam Judul Prometeus Setelah itu meminjam Judul Teseus-Perseus Setelah itu meminjam Judul Kisah persefone (support = 2) Setelah itu meminjam Judul Pertempuran para Titan Setelah itu meminjam Judul Teseus-Perseus (support = 2) Setelah itu meminjam Judul Pertempuran para Titan Setelah itu meminjam Judul Teseus-Perseus Setelah itu meminjam Judul Kisah persefone (support = 2) Setelah itu meminjam Judul Pertempuran para Titan Setelah itu meminjam Judul Kisah persefone (support = 2)
Judul Pengantar Tata Hukum Indonesia (support = 4)	

Figure 8. Sequential Pattern Rules in the form of codes (left) and bahasa Indonesia (right)



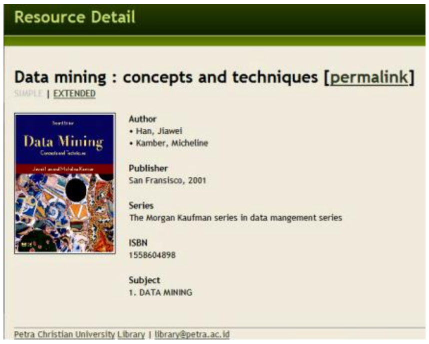
```
⊟1-29366 Judul Schaum's Outline of theory and Problems of Fluid Mechanics and Hydraulics
   1-58281 Judul Schaum's outline of theory and problems of fluid mechanics and hydraulics
 1-29369 Judul Schaum's outline of English grammar
 1-2961 Judul Architecture
 1-29606 Judul Segi-segi hukum hak milik intelektual
 1-29614 Judul Seismic design of reinforced concrete and masonry buildings
 1-2964 Judul Architecture
⊟1-29697 Judul Sejarah revolusi kemerdekaan daerah istimewa aceh
   1-1293 Judul Adat istiadat daerah propinsi Daerah Istimewa Aceh
 1-29731 Judul Seksualitas pornografi pernikahan
⊟1-29822 Judul Selling
  ⊟1-69981 Judul Principles of services marketing
     1-11182 Judul Fundamentals of marketing
   1-11182 Judul Fundamentals of marketing
  ⊟1-46712 Judul Promosi penjualan
     1-11182 Judul Fundamentals of marketing
    ⊟1-69981 Judul Principles of services marketing
       1-11182 Judul Fundamentals of marketing
 1-29830 Judul Semantics
 1-29832 Judul Semantics
 1-29937 Judul Seni berkomunikasi
```

Figure 9. Sequential Pattern Tree

All sequential patterns will be used as the basis for the suggestion that displayed on the web page of the book you looking for in the website of PCU library. There are two kinds of suggestion that are produced, namely:

- A list of books frequently borrowed along with the book that is being viewed, under the header "People who borrowed this book also borrowed ..."
- A list of frequently borrowed book after borrowing a book that is being viewed, under the header " People who borrowed this book then borrowed ..."

The original webpage can be seen in Figure 10. The alterations can be seen on Figure 11.



Resource Detail

Data mining : concepts and techniques [permalink]

SIMPLE | EXTENDED

Author
• Han, Jiawei
• Kamber, Micheline

Publisher
San Fransisco, 2001

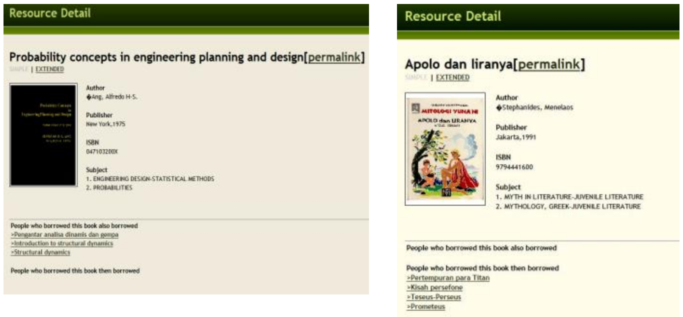
Series
The Morgan Kaufman series in data mangement series

ISBN
1558604898

Subject
1. DATA MINING

Petra Christian University Library | library@petra.ac.id

Figure 10. The original webpage of PCU library books searching system.



Resource Detail

Probability concepts in engineering planning and design[permalink]

SIMPLE | EXTENDED

Author
◆Ang, Alfredo H-S.

Publisher
New York,1975

ISBN
047103200X

Subject
1. ENGINEERING DESIGN-STATISTICAL METHODS
2. PROBABILITIES

People who borrowed this book also borrowed
>Pengantar analisa dinamis dan gempa
>Introduction to structural dynamics
>Structural dynamics

People who borrowed this book then borrowed

Resource Detail

Apolo dan liranya[permalink]

SIMPLE | EXTENDED

Author
◆Stephanides, Menelaos

Publisher
Jakarta,1991

ISBN
9794441600

Subject
1. MYTH IN LITERATURE-JUVENILE LITERATURE
2. MYTHOLOGY, GREEK-JUVENILE LITERATURE

People who borrowed this book also borrowed

People who borrowed this book then borrowed
>Pertempuran para Titan
>Kisah persefone
>Teseus-Perseus
>Prometeus

Figure 11. Example of the alteration 1 (left) and 2 (right) of PCU library books searching system

IV. Testing Results

A To test the application, we perform two kinds of tests, namely:

- Testing the speed of the CloSpan process to generate sequential patterns. The result is shown in Figure 12.
- Tests on the total rules generated. The result is shown in Figure 13.

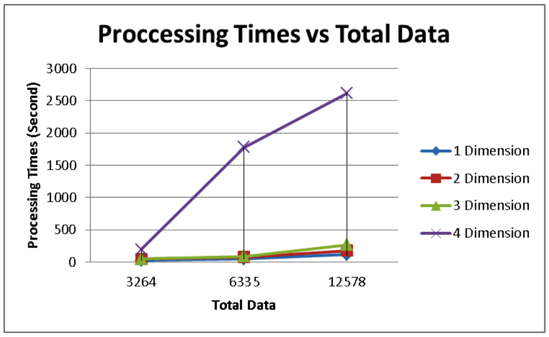


Figure 12. The result of processing time test

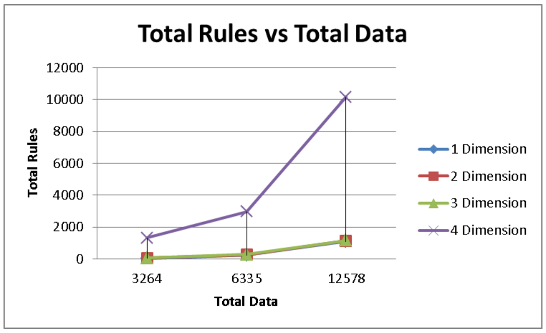


Figure 13. The result of total rules generated test

In addition to these two tests, testing was also conducted in the form of questionnaires to some people with an interest in this application, such as the head of PCU library along with his staff, and also some students who often use of this library service. The questionnaires results are shown in Table 1.

TABLE I. THE QUESTIONNAIRES RESULTS

No	Job	Criteria				
		A	*B*	*C*	*D*	*E*
1	Head of PCU Library	3	3	4	4	3
2	Collection Sector Supervisor	3	4	4	4	3
3	Junior Programmer	4	4	5	4	4
4	Customer Service Supervisor	4	5	5	5	4
5	Books Processing Staff	5	5	4	4	5
6	Automation Processing Staff	3	3	4	4	2
7	PCU Informatics Dept. Lecturer	4	4	4	4	3
8	PCU Informatics Dept. Lecturer	5	4	4	5	5
9	PCU Informatics Dept. Student	5	4	5	5	4
10	PCU Informatics Dept. Student	4	3	4	4	4
Total		40	39	43	43	37
Mean (Percentage)		80%	78%	86%	86%	74%

Scoring: 1 → Very Bad to 5 → Very Good

The criteria:

A = Ease of use of the application.

B = Interface design (appearance) of the application.

C = The accuracy of the information produced.

D = Applications can meet the needs of the library.

E = The use of language in the application.

V. Conclusion

This application has been designed and implemented correctly. This can be seen in the results of application testing. And also the results of questionnaires from library staffs and prospective users generate good points.

References

[1] G. Dong, and J. Pei, Sequence Data Mining, Springer Science + Business Media, 2007

[2] G. S. Budhi, A. Handojo, and S. G. Sutrisno, “Book Loan Recommendation System for Petra Christian University Library using PrefixSpan and Generalized Sequential Pattern Algorithm”, Proc. The 2nd Makassar Int. Conf. on Electrical Engineering and Informatics (MICEEI’10), pp. 136–143, Makassar, Indonesia, October 2010

[3] J. Han, M. Kamber, and J. Pei, Data mining: Concepts and techniques, 3rd Edition, Elsevier Inc, 2012

[4] R. Agrawal, and R. Srikant, "Mining sequential patterns", Proc. 1995 Int. Conf. Data Engineering (ICDE’95), pp. 3–14, Taipei, Taiwan, March 1995

[5] X. Yan, J. Han, and R. Afshar, “CloSpan: Mining closed sequential patterns in large datasets”, Proc. 2003 SIAM Int. Conf. Data Mining (SDM’03), pp. 166–177, San Fransisco, CA, May 2003