International Journal of Informatics and Information System ISSN 2579-7069
Vol. 2, No. 1, March 2019, pp. 17-23

Analysis of Sequential Book Loan Data Pattern Using Generalized Sequential Pattern (GSP) Algorithm

Tri Astuti [a,1*], Lisdya Anggraini [a,2]

[a] Informatics Engineering Program, Faculty of Computer Science, Amikom Purwokerto University
[1] tri_astuti@amikompurwokerto.ac.id; [2] lupherslisdya08@gmail.com
* corresponding author

Abstract

As a center for learning and information services, STMIK Amikom Purwokerto Library is a source of learning and a source of intellectual activity that is very important for the entire academic community in supporting the achievement of the college Tridharma program. Book lending transaction data, can produce information that is important as supporting decision making when further analyzed. One useful information is that it can provide information in the form of user behavior patterns in borrowing books that are used to maintain the availability of related book stocks to be balanced. This study uses the Generalized Sequential Pattern (GSP) algorithm, which can be used to determine the behavior patterns of users in each transaction and can show relationships or associations between books, both requested simultaneously and sequentially. From the calculations that have been done, 295 frequent sequences are consisting of 3 sequence patterns that are formed from the minimum support of 0.53% or the minimum number of books borrowed, namely 2 books. Three book items have very strong linkages in book lending transactions, namely book code 6690, 2026, and 8131.

Keywords: Data Mining, Association Rules, Apriori Algorithm, Minimal Support, Confidence.

1. Introduction

The library is a venue, a building reserved for the maintenance and use of books and so on, can also be interpreted as a collection of books, magazines, and other literature materials stored for reading, learning, talk about[1]. In addition to serving the collection of writings, prints, and/or records, currently, the library is considered the resource of information that is the mobilizer of an institution[2][3].

Book-Lending transaction Data can produce valuable information as support for decision making when analyzed further. One useful information is to provide information in the form of student behavior patterns in borrowing books, providing information about related books, keeping the stock availability of related books to be balanced, arrangements The placement of books related to bookshelves, and many other beneficial strategies that can be used as supporting in making decisions related to the management of book stock in the library STMIK Amikom Purwokerto.

Further analysis of book loan data can be done by implementing data mining. Data mining is an essential process of extracting information or patterns in large databases [4]. The application of data mining in book-lending data in the library of Stmik Amikom Purwokerto is expected to be used as a support for decision making by looking at the book lending pattern.

There has been previous research on the processing of book-lending data in the library of Stmik Amikom Purwokerto by combining the apriori algorithm and the fp-growth algorithm conducted by [5] obtained 5 association rules from Minimum support 0.01 (1%) and minimum confidence 0.5 (50%) [6].

While the research conducted by [7] is about the development of market basket analysis application in supermarkets using the algorithm of a generalized sequential pattern. The results of the analysis can be used as a strategy for running a business, such as a layout recommendation of goods and maintaining the availability of product stock related to being balanced. From the test application conducted from transaction data period August 10, S. D August 25th, obtained the information in the form of rules from minimum support 50% resulted in rules 18, number of transactions 10, number of customers 4, number of item 25 and execution time 00:00:00:604.

2. Literature Review

2.1. Generalized Sequential Pattern (GSP) algorithm

The GSP algorithm is used in the mining sequence and is useful for solving many Mining sequence issues based on a priori algorithm [8]. The primary function of the GSP algorithm is to find a pattern. [9]

Data extracting sequential patterns try to find relationships between occurrences of a sequential event in order to search for a specific sequence of events. In other words, a sequential pattern excavation aims to find a frequently occurring sequence to describe data or data that predicts a future or periodic pattern excavation[10].

Generalized Sequential Patterns (GSP), can obtain data information about the frequently borrowed books (the rules) and frequently borrowed books sequentially (sequential pattern Rules) by the same borrower. With the GSP algorithm, both kinds of information will be obtained simultaneously in one process [11][12].

2.2. Sequential Pattern Mining Framework (SPMF)

Sequential Pattern Mining Framework (SPMF) is a library function written in the Java programming language to handle data Mining tasks with the Open-source GPL v3 license[13]. For the result of the input file, there is a Format output file defined as a text file. Each row is a frequent sequential pattern. Each item of a sequential pattern is a positive integer and an item of the same itemset in a sequence separated by one space. The value "-1" indicates the end of an itemset. The value "-2" indicates the end of a sequence (this appears at the end of each line). On each row, the sequential pattern is indicated first. Then, the keyword "#SUP:" appears followed by an integer that demonstrates the support of Dari pattern as some sequences[14][15].

3. Method

This research is conducted in the library of STMIK AMIKOM Purwokerto, located at Jl. Letjend. Pol. Sumarto front NES Purwoketo Watumas. The research uses the lending data from 20 March 2017 to 16 August 2017.

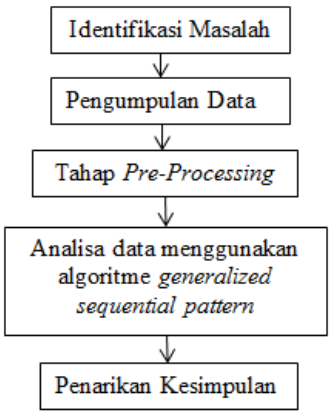

Fig 1. Research Framework

The research flows in this study are as follows:

3.1. Problem Identification

The process of identification of the problem is done as an effort to know the problem in Amikom library and using the new method so it can be determined that several points analyze algorithm performance in book lending

3.2. Data Collection

In this study, the primary data used by book-lending data can be from the library of STMIK AMIKOM Purwokerto. The Data consisted of 1425 Records, which were started from 20 March 2017 – 16 August 2017. Also, data of information about the library is a library profile, graphs, and interviews to obtain information about the library.

3.3. Pre-Processing Stage

At this stage, the data selection process to obtain data is clean and ready to be used as research material. These stages include data selection, data sanitization, Data transformation. This research is done by analyzing the type of book in

each transaction, not the number of books in each transaction and to find the relation between books. Therefore, the result of data ready to be filtered back leaves a top record when in one advanced a name the same book.

3.4. Data Analysis Using Generalized Sequential Pattern Algorithm

The algorithm that the author uses in this Study is a generalized sequential pattern algorithm.

3.5. Conclusion

At this stage, the authors conclude from the results of the study that has been done produced a book lending pattern that is formed from the Use of the generalized sequential pattern algorithm.

4. Results and Discussion

To determine the sequential pattern with the GSP algorithm is done with the help of the SPMF application. This application reads input in the form of data in sequential form. Before the data is in the process through the SPMF application, it is initialized to the book code.

Table 1. Initials Code Book

Initials	Book Code
1	7398
2	6690
3	10495
4	10559
5	9122
...	...
129	4755
130	4753
131	77
132	7189
129	4755
...	...
631	8129

Analisa generated by using the GSP algorithm with a minimum of 0.53% support in obtaining results is using the execution time GSP algorithm required to form a sequential pattern of the SPMF software generating 302 frequent sequences. From the calculated result, the minimum support used is 0.53%, and the minimum confidence is 50% obtained the following results.

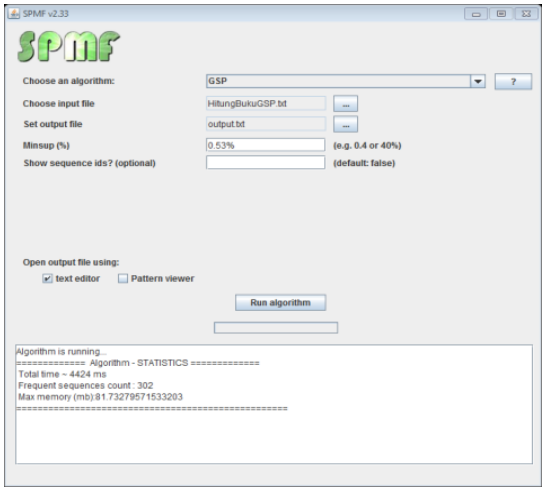

Fig. 2 SPMF Data input display

The following results of the GSP algorithm output using SPMF software.

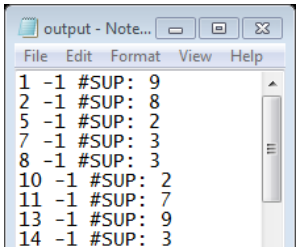

Fig. 3 Display output DATA SPMF 1 – Sequences

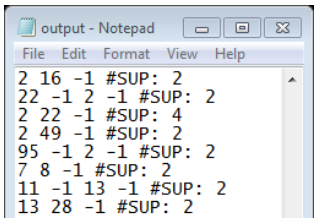

Fig. 4 Display Output Data Spmf 2 – Sequences

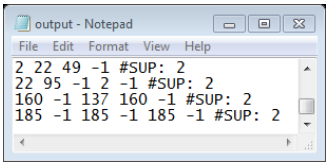

Fig. 5 Display Output Data SPMF 3 – Sequences.

After the data analysis using the SPMF software is complete, the process of datasets is then used by Microsoft excel 2010. This stage aims to obtain the Association rules using the generalized Sequential Pattern algorithm. From the calculated result, the minimum support used is 0.53%, and the minimum confidence is 50% obtained the following results.

4.1. The Generate Frequent Itemset Process

The generate frequent itemset process is the process of forming a candidate itemset and its support to obtain a frequent itemset that satisfies the minimum support using the GSP algorithm.

a. Calculation minimum of support for the appearance of each item.

Table 2. The 1st Frequent itemset

No	Initial	*Item*	*Support*
1	1	7398	9
2	2	6690	8
3	5	9122	2
4	7	2102	3
5	8	1140	3
...	...	...	...
241	576	2458	2
242	589	6552	2
243	591	1790	2
244	603	927	2

From the result of the calculation of the obtained 244 items that meet The minimum support, 0.53% or the number of books in the book is 2 or more.

b. The search candidate Itemset 2nd non-sequential or sequential is done by searching for the Item column. For candidate Itemset non-sequential when finding candidate [1.2] and [2.1], then considered the same. The sequential candidate for a candidate [1 → 2] and [2 → 1] is distinct.

Table 3. Non-Sequential 2nd Frequent itemset

No	Inisial	*Item* Non-Sekuensial	*Support*
1	2 16	6690 8984	2
2	2 22	6690 2026	4
3	2 49	6690 8131	2
4	7 8	2102 1140	2
...	...	...	...
28	438 476	5346 5392	2
29	454 591	4472 1790	2
30	522 523	2542 6146	2

The non-sequential Items field is a borrowed item at the same time, while for the sequential Item field The item is borrowed sequentially.

Table 4. Sequential 2nd Frequent itemset

No	Inisial	*Item* Sekuensial	*Support*
1	22 -> 2	2026 -> 6690	2
2	95 -> 2	7011 -> 6690	2
3	11 -> 13	9967 -> 6389	2
4	16 -> 97	8984 -> 5642	2
...	...	...	...
15	160 -> 163	10087 -> 1639	2
16	164 -> 302	1397 -> 10581	2

Table 3 and Table 4 are the 2nd candidate Itemset that meets the minimum support. The results of the non-sequential frequent itemset result are received 30 non-sequential patterns (in the same loan) with the highest support

value of 5 with the initials 22 49 (2026 8131). Moreover, the results of the frequent itemset to-2 sequentially obtained 16 sequential patterns that are borrowed sequentially by users with all the results of its support value 2.

c. The establishment of the 3rd candidate Itemset is depicted by joining the result of the frequent itemset to-1 with the results of the 2nd frequent itemset — further calculations for the number of his support.

Tabel 5. Non-Sequential 3rd Frequent itemset

No	Inisial	*Item* Non-Sekuensial	*Support*
1	2 22 49	6690 2026 8131	2

Tabel 6. Sequential 3rd Frequent itemset

No	Inisial	*Item* Sekuensial	*Support*
1	22 95 -> 2	2026 7011 → 6690	2

From the result of the calculation, obtained 1 non-sequential pattern and 1 sequential pattern with the same support value of 2.

d. The iteration of this GSP algorithm will stop if it can no longer be found next candidate Itemset that can be formed. Table 5 and Table 6 are examples of the last result of frequent itemset that can result from transaction data.

4.2. The Generate Rules

The generate rule process serves to generate rules by processing the data in the frequent itemset table that has been generated in the generate frequent items process.

Table 7. Results Generated Rules

Freq Item	*Rule*	*Confidence*
2 22	(2) → (2 22)	4/8 = 50%
7 8	(7) → (7 8)	2/3 = 66,6%
	(8) → (7 8)	2/3 = 66,6%
13 28	(28) → (13 28)	2/4 = 50%
22 197	(197) → (22 197)	2/2 = 100%
55 159	(55) → (55 159)	2/4 = 50%
95 96	(95) → (95 96)	3/5 = 60%
	(96) → (95 96)	3/4 = 75%
115 116	(115) → (115 116)	2/2 = 100%
	(116) → (115 116)	2/2 = 100%
130 185	(185) → (130 185)	2/4 = 50%
137 160	(160) → (137 160)	2/3 = 66,6%
163 496	(496) → (163 496)	2/2 = 100%
181 603	(181) → (181 603)	2/3 = 66,6%
	(603) → (181 603)	2/3 = 66,6%
200 270	(270) → (200 270)	2/3 = 66,6%
232 233	(232) → (232 233)	2/3 = 66,6%
	(233) → (232 233)	2/2 = 100%
438 476	(476) → (438 476)	2/4 = 50%
454 591	(454) → (454 591)	2/4 = 50%
	(591) → (454 591)	2/2 = 100%
522 523	(522) → (522 523)	2/2 = 100%
	(523) → (522 523)	2/3= 66,6%
2 22 49	(2, 22) → (2 22 49)	2/4 = 50%
	(2, 49) → (2 22 49)	2/2= 100%

From the calculation of the above, comes the result with the best rule 25 rule derived from 17 frequent itemset consisting of 7 rules that meet the value of confidence 50%, 1 Rule that meets confidence value 60%, 8 rules that meet confidence value 66.6%, 1 rule that meets confidence value 75%, and 8 rules that meet confidence value of 100%. DEngan minimum support 0,53 and minimum confidence 50% or a minimum number of books borrowed 2.

So the book lending pattern in STMIK Amikom Purwokerto Library can be made by looking at the book relationship that can be seen in table 4.12. Here are some examples of book lending patterns from calculations conducted using the generalized sequential pattern algorithm, namely:

a. If users borrow Buku "multimedia konSep & Application in education" then will borrow the book "multimedia konSep & Application in education" and "multimedia DigitaL base theory + development".
b. If a user borrows a Buku "Rational Rose untofan object-oriented modeling" then it will borrow the book "Rational Rosefor themodeling of object-oriented" and "software-oriented engineering with the USDP (Unified SSoftware development Process") method.
c. If users lend a book "Multimedia konSep & Application in education" with the book "method of research qualitative Kand R & D" then will borrow the book "Multimedia konSep & Application in Education", "Digita multimediabase theory + development" and the book "method Qualitative Quantitative research and R & D ".

5. Conclusion

B theamount of rule generated is influenced by the large number of transaction data, borrowed items simultaneously and the number of users. In addition, the pattern of the book-Lending behavior by users in different times can reappear at a later time.

From PErcount that has been done obtained 295 frequent sequence consisting of 3 sequence patterns formed from Minimum support 0.53% or the minimum number of books borrowed is 2 books and minimum confidence 50%. For the best rule of 25 rule. One of the rules with the highest confidence level (confidence) of 100% i.e. If a user is borrowing a book with initials 2 (6690 □ Multimedia Concept & application in education) with initials 49 (8131 □ Qualitative Quantitative research method and R & D) then it will borrow a book with initials 2 (6690 □ Multimedia Concept & Applications in Education), initials 22 (2026 □ Digital Multimedia base theory + development) and initials 49 (8131 □ methods of research qualitative Kand R & D).

References

[1] P. N. Tan, M. Stenbach, and V. Kumar, “Introduction to Data Mining”. Boston: Pearson Education, 2006.
[2] F. Goronescu, “Data Mining: Concepts, Models, and Techniques.” Verlag Berlin Heidelberg: Springer. 2011.
[3] I. H. Witten, E. Frank, and M. A. Hall. “Data Mining: Practical Mchine Learning Tools and Techniques, 3rd ed”. Burlington, MA: Morgan Kaufmann, 2011.
[4] Berkhin, P., 2002. A Survey of Clustering Data Mining Techniques, Technical Report. Accrue Software, San Jose, CA.
[5] Chen, C.-H., Hong, T.-P., Tseng, V.S., Lee, C.-S., 2009. A genetic-fuzzy mining approach for items with multiple minimum supports. Soft Computing – A Fusion of Foundations, Methodologies and Applications 13 (5), 521–533.
[6] Chen, S.-S., Huang, T.C.-K., Lin, Z.-M., 2011. New and efficient knowledge discovery of partial periodic patterns with multiple minimum supports. Journal of Systems and Software 84 (10), 1638–1651.
[7] Chiang, D.-A., Wang, Y.-H., Chen, S.-P., 2010. Analysis on repeat-buying patterns. Knowledge-Based Systems 23 (8), 757–768.
[8] Chu, C.-J., Tseng, V.S., Liang, T., 2008. An efficient algorithm for mining temporal high utility itemsets from data streams. Journal of Systems and Software 81 (7), 1105–1117.
[9] Chun, S., Park, Y., 2006. A new hybrid data mining technique using a regression case based reasoning: application to financial forecasting. Expert Systems with Applications 31, 329–336.
[10] Ezeife, C., Lu, Y., 2005. Mining web log sequential patterns with position coded preorder linked wap-tree. Data Mining and Knowledge Discovery 10, 5–38.
[11] Han, J., Pei, J., Yin, Y., Mao, R., 2004. Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Mining and Knowledge Discovery 8, 53–87.
[12] Han, J., Cheng, H., Xin, D., Yan, X., 2007. Frequent pattern mining: current status and future directions. Data Mining and Knowledge Discovery 15, 55–86.
[13] Hu, Y.-H., Chen, Y.-L., 2006. Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism. Decision Support Systems 42, 1–24.
[14] Hu, Y.-H., Wu, F., Liao, Y.-C., 2010. Sequential pattern mining with multiple minimum supports: a tree based approach. In: The 2nd International Conference on Software Engineering and Data Mining, Chengdu, China.
[15] Kim, E., Kim, W., Lee, Y., 2003. Combination of multiple classifiers for the customer’s purchase behavior prediction. Decision Support Systems 34, 167–175.