Mining fuzzy time interval sequential pattern on event log data using FP-Growth-Prefix-Span algorithms

Imam Mukhlash' , M. Sidratul Muntaha A. M. A.' , Mohammad Iqbal' , Ahmad Saikhu' , and Riyanarto Sarno'

Citation: AIP Conference Proceedings **1746**, 020065 (2016); doi: 10.1063/1.4953990
View online: http://dx.doi.org/10.1063/1.4953990
View Table of Contents: http://aip.scitation.org/toc/apc/1746/1
Published by the American Institute of Physics

Mining Fuzzy Time Interval Sequential Pattern on Event Log Data using FP-Growth-Prefix-Span Algorithms

Imam Mukhlash[1, a)], M. Sidratul Muntaha A.M.A[1,b)], Mohammad Iqbal[1, c)], Ahmad Saikhu[2, d)], Riyanarto Sarno[2, e)]

[1]*Department of Mathematics, Faculty of Sciences, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[2]*Department of Informatics, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*

[a)]Corresponding author: imamm@matematika.its.ac.id
[b)]m.sidratul.muntaha92@gmail.com
[c)]iqbalmohammad.math@gmail.com
[d)]saikhu@cs.its.ac.id
[e)]riyanarto@if.ac.id

Abstract. Rapid technological developments caused the increasing number of computerized data processing. With the increasing complexity of business processes, business process management technologies such as ERP (Enterprise Resource Planning) are increasingly being used. This resulted in the availability of data more abundant so that excavation and search information from the dataset will be a valuable knowledge. In this paper, we have done the process mining to obtain an interesting pattern of event log data. In this research, data mining method that we are used is the sequential pattern mining algorithm using FP-Growth-Prefix Span. In addition, we are also used the fuzzy approach to handle the time interval of the analyzed data, so that the sequential pattern that produced become fuzzy time-interval sequential pattern. The application of these methods in a business processes that produce fuzzy time interval sequential pattern. From the analysis, the result shown that there is a minimum effect on the pattern of the resulting support. Furthermore, the results of the analysis can be used as consideration in the analysis of business processes.

Keywords: Process Mining, Business Processes, FP-Growth, PrefixSpan, Fuzzy time-interval sequential pattern

INTRODUCTION

Along with the development of computer technology is increasingly rapidly over the past few years, business process management technology is also increasingly being used [1], [2]. It was also in line with the number of existing business processes. With the business process management technology, a company can build and update any information in the business process quickly included in the repository model of the process so that every service provided by these companies can change rapidly.

Increasing use of business process management technology can be seen from the many companies that work with automation performance for their company. Enterprise Resource Planning (ERP) is a popular example of the use of business process management technology. This technology is applied by many large companies. A large company would have hundreds or even thousands of business processes. Discover and analyze the similarity of the collection process owned businesses will be very useful for the company. For example, when several companies have different

The 2016 Conference on Fundamental and Applied Science for Advanced Technology (ConFAST 2016)
AIP Conf. Proc. 1746, 020065-1–020065-6; doi: 10.1063/1.4953990
Published by AIP Publishing. 978-0-7354-1403-7/$30.00

business processes to join, it must be known to the similarity of business processes across the enterprise so that it can know where the business processes of a company can be incorporated. In the end, processing and in-depth analysis of the collection of existing business process models into a valuable knowledge. Processing and analysis can be done by applying data mining [2].

Data mining is the process of extracting important information or patterns in the databases. The application of data mining in the business process model is expected to be able to gather information and analyze the results to improve the efficiency, notice the trend and determine the standardization of business processes. Based on previous studies [2], it is said that in measuring the similarity value of business processes used integration metric between semantic and structural similarity. The integration of semantic and structural similarity is done by weighting the value. In this study, a clustering technique based on the distance between two similar entities using graph partition method was used. This method is usually used when the cluster object is difficult to be represented into mathematics form. The distance which is used in the clustering process based on the similarity of each analyzed model and business processes sets with high similarity would be one cluster [3]. Moreover, there are several researches about business process analysis by using process mining, i.e., [4]. In this research, we use data mining technique to analyze business process sequential through event log data.

One of task in data mining can be used is sequential pattern mining. Sequential pattern mining is useful to get pattern on event sequential that applied when the usage software recorded on an event log. From those patterns can be used to find the user trends in utilizing software for fraud detection [5], next event transaction and others. One of the successes of the business process is the regularity of the transaction process. The transactions or events pattern that occurs in the business process model is usually portrayed in a sequential pattern. In this research, we use sequential pattern mining as data mining method by using FP-Growth-Prefix-Span algorithm because it has better performance than other algorithm such as GSP and other Apriori-like algorithms, Free-Span, or SPADE [6][7]. Sequential pattern indicates that the transaction in the business process usually occurs serially over the time. We need to further analyze the effect of the length of time between the occurrences of the events in the event log, whether short, medium or long. Therefore, we utilize the fuzzy time approach to analyze this data and use fuzzy time interval-PrefixSpan algorithm in [7] to find the fuzzy time interval sequential patterns on event log data.

BASIC THEORY AND RELATED RESEARCH

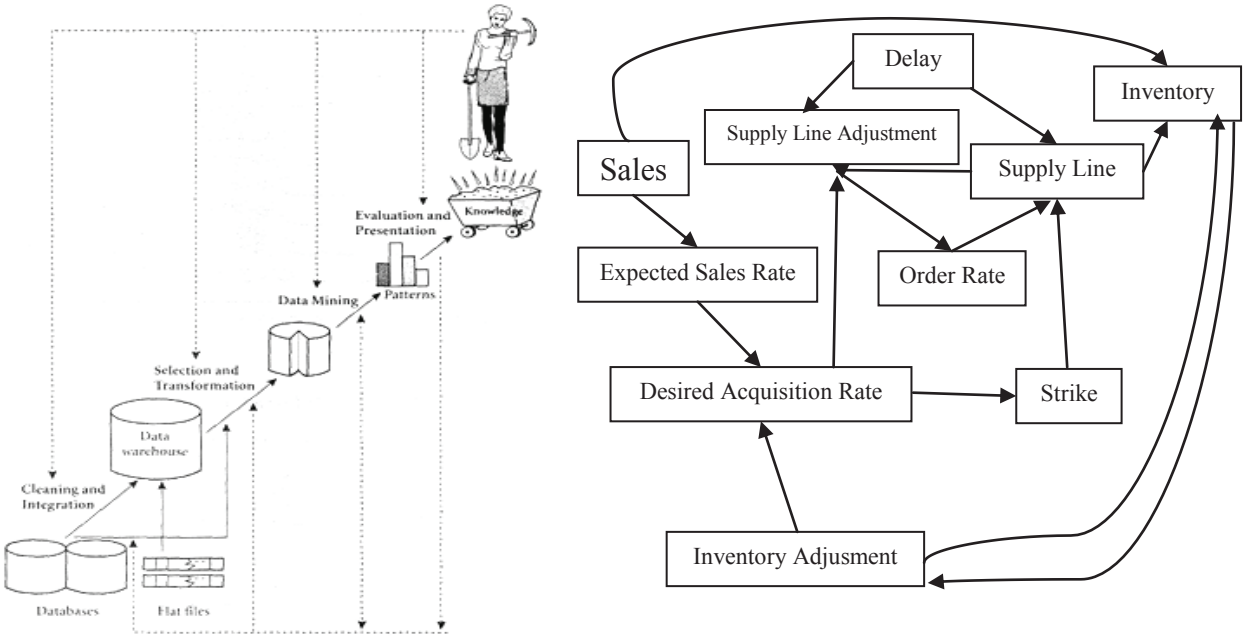
Data Mining and Knowledge Discovery in Database (KDD)

Data Mining is the extraction process or patterns of important information in large databases [8]. Data mining is one step in the process of Knowledge Discovery in Databases (KDD) to find useful patterns. Data Mining is also defined as a process that uses a variety of data analysis tools to find patterns and relationships of data that can be used to make exact predictions. Knowledge Discovery in Databases (KDD) is the process of finding useful information and patterns that exist in the data. KDD is a process that consists of a series of sequential, iterative process, and data mining is one step in the KDD process [8]. The stages sequentially KDD process can be seen in Figure 1 (a).

Business Process, Mining Process and Sequential Pattern

Business process consists of a series of activities were carried out with the coordination of organizational and technical environment. These activities together build business goals. Every business process defined by the organization, but can interact with business processes implemented by other organizations. Business processes have specific input and output, resources, and has activities in a certain order [9].

A business process model consists of a series of activities models and constraint execution between all of them. The example of business process is a real case in the operational business of companies, consisting of examples of activities. Each business process model acts as the blueprint for a series of example of business processes, and each model activities act as a blueprint for a series of examples of activities. Business process consists of various activities in accordance with business objectives. Such activities could be a system of activities, activities of user interaction such as sending packets to a business associate, or manual activities. Manual activities are not supported by the information system. An example of business process model is shown in Figure 1 (b).



(a) (b)

FIGURE 1. (a) KDD Process [8] and an example of business process model [9].

Process mining is used to analyze the existing business processes based on the event log. On the process mining, the observations were made on the business processes that have been computerized. With this way, the new process structure can be found which is previously unrecognized happening. Based on a frequency of information flow that occurs and consistent cycle, it can be seen whether the business processes implemented by the information system in accordance with the guidelines of the organization or vice versa.

Sequential pattern is a list of the order set items. In a business process, a sequential pattern defined as an instance B in an active process p after the completion of instance A at p. The transaction pattern or events that occur in the business process model is usually portrayed in a sequential pattern. Sequential pattern indicates that the transaction usually occurs serially over time. Therefore, sequential pattern analysis of business processes based on chronological order or sequential of the occurrence of a transaction. Given a minimum support is positive integer as the support threshold and a sequence α-called sequential pattern in a sequence database S if $\text{support}_s(\alpha) \geq \text{minimum support}$.

Mining Fuzzy Time Interval Sequential Patterns using Prefix-Span Algorithm

Yen-Liang, et. al. have developed Prefix-Span based algorithm to find sequential pattern with fuzzy time interval, namely FTI-Prefix-Span algorithm with steps as follow[7]:

1. Determine linguistic term from time interval of linguistic variable then find the degree or value of membership through a membership function.
2. Build fuzzy time interval sequence database
3. Find all frequent items in fuzzy time interval sequence database, so that discovered sequential pattern of fuzzy time interval length-1. Then, count frequencies of each item in fuzzy time interval sequence database. All items with $support$ value $\geq minimum\ support$ are elements from sequential pattern length-1. Sequential pattern length-1 obtained can be considered as prefix.
4. For search space by using the prefix obtained in step 1. The prefixes will continually changes as the iteration process of finding a sequential pattern length-k with$k > 1$.
5. To search space prefix 1, get subsets sequential pattern using projected sequence database fuzzy time interval. The database is projected formed by taking the suffix of sequence databases based prefix obtained in the previous step. Then, calculate the degree of membership of each item for each linguistic term in the database

projected sequential fuzzy time interval. Use the definition 5 to find support for each linguistic term. Support with linguistic terms greater than or equal to the minimum support is a member of a fuzzy time interval sequential pattern length-2. Then, make a sequential pattern obtained as a new prefix for the next search. The next projected databases established by the new prefix are generated. Next, do the search process is repeated in this subset.

6. Do a search for other prefixes sequential pattern (sequential pattern length-1) and a search process as in step 3.

RESULT AND ANALYSIS

In mining fuzzy time intervals sequential pattern required on business processes analysis and design are good, ranging from system analysis to software design. Explanations related to it will be described as follows.

Software Analysis and Design

The purpose of modeling analysis is to explain the systematic analysis of the passage of the application to be made. The following image is given use case and activity diagrams for extracting application fuzzy time interval sequential pattern as seen in Figure 2.

Database design made when the used data is a collection of various interconnected tables and forms a large database. In this study, we made simple database design because the used data only an event log table. Then, we perform the preprocessing data design that provides an explanation of the process that we made from the initial data, thus this data can be used in the data mining process. In this research, the used data from an excel file with only one sheet or table consist two attributes such as task (events/instances) and time-stamp. This data is about the company's business process which is generated as previously event log but has normalized [10]. The task consists of eight activities of business process such as Register, Analyze Defect, Repair (Complex), Test Repair, Inform User, Archive Repair, Repair(Simple), and Restart Repair. Number of this record data is 11.854 records. Furthermore, we do some data processes such as data cleaning and data transformation so that sequential time interval database formed as much as 1.103 records. On the data mining process design, we perform mining sequential pattern using mining fuzzy time interval sequential pattern algorithm. The steps of data mining process are import data, mining sequential pattern mining length-1, mining fuzzy time interval sequential pattern length-2 and length more than two. Linguistic terms that will be used for linguistic variable time interval are short, middle and long which is defined by following membership function [7],

$$\mu_{short}(ti_j) = \begin{cases} 1 & ti_j \le 2 \\ \dfrac{15 - ti_j}{13} & 2 < ti_j < 15 \\ 0 & ti_j \ge 15 \end{cases} \quad (1)$$

$$\mu_{middle}(ti_j) = \begin{cases} 0 & ti_j \le 2 \quad or \quad ti_j \ge 28 \\ \dfrac{ti_j - 2}{13} & 2 < ti_j \le 15 \\ \dfrac{28 - ti_j}{13} & 15 < ti_j < 28 \end{cases} \quad (2)$$

$$\mu_{long}(ti_j) = \begin{cases} 0 & ti_j \le 15 \\ \dfrac{ti_j - 15}{13} & 15 < ti_j < 28 \\ 1 & ti_j \ge 28 \end{cases} \quad (3)$$

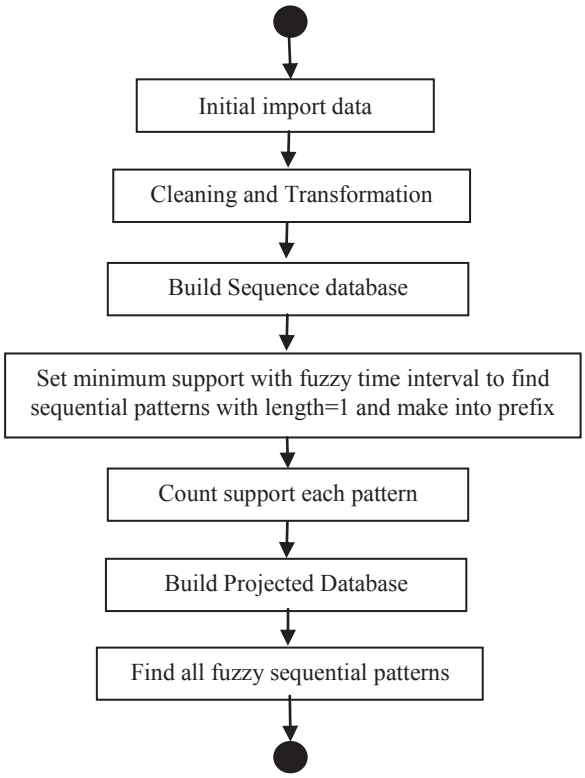


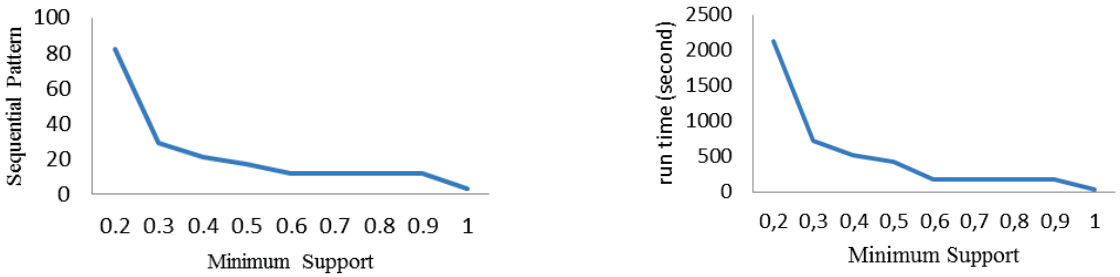
FIGURE 2. Generally Framework of Mining Fuzzy Time Interval Sequential Pattern

The result of mining sequential pattern for business process using system that we built can be seen in Table 1 when given minimum support is 50%. From Table 1, it describes that all attributes of business process on event log data generates only 20 attributes as shown in Table 1 with a given minimum support is 0.5 or 50% from the number of datasets. Moreover, it can be seen that Register, AnalyzeDefect and Register.short.AnalyzeDefect are attributes which are strongly recommend become factors in the analysis of business process because it have support value equal to 1. Figure 3 (a) shows the relation between minimum support and number of generated sequential pattern. This graph explains that the number of generated sequential patterns is decrease when the minimum support value is increase because the support value of sequential pattern is smaller than the given minimum support values and it means that mining sequential pattern mining can be used for analysis business process on event log data. Figure 3 (b) mean about the less time it takes to find a fuzzy time interval sequential pattern when the minimum support value is increase is due to there are a lot of sequential pattern with the support value is below the given minimum support, thus generated sequential pattern is decrease. When sequential pattern length-k produced is low then the searching process of sequential pattern length-(k + 1) will be faster. It means that by using fuzzy time interval can be optimized the performance of mining sequential pattern mining on event log data.

Table 1. Generated Sequential Pattern when minimum support = 0.5

Sequential Pattern	*Support*	Sequential Pattern	*Support*
Register	1	Register,short,AnalyzeDefect,short, AnalyzeDefect	0,57
AnalyzeDefect	1	Register,short,AnalyzeDefect, long,TestRepair	0,910
Repair(Complex)	0,596	Register,short,AnalyzeDefect, long, InformUser	0,949
TestRepair	0,998	Register,short,AnalyzeDefect, long, ArchieveRepair	0,905
InformUser	0,998	Register,short,AnalyzeDefect,short, AnalyzeDefect,long,TestRepair	0,743
ArchieveRepair	0,905	Register,short,AnalyzeDefect,short, AnalyzeDefect,long,TestRepair,long,InformUser	0,776
Register,short,AnalyzeDefect	1	Register,short,AnalyzeDefect,short, AnalyzeDefect,long,TestRepair,long,ArchieveRepair	0,902

Register,long,TestRepair	0,909	Register,short,AnalyzeDefect,short, AnalyzeDefect	0,57
Register,long,InformUser	0,949	Register,short,AnalyzeDefect, long,TestRepair	0,910
Register,long,ArchieveRepair	0,905	Register,short,AnalyzeDefect, long, InformUser	0,949



(a) (b)

FIGURE 3. (a) Relation between minimum support and sequential pattern and (b) Relation between minimum support and run time

CONCLUSION

In this paper, we have applied fuzzy time interval Prefix-Span based algorithm to discover sequential patterns on event log data. The steps taken are cleaning and transformation for initial event log data, build sequence database, set and count fuzzy support parameter, build projected database, and find all sequential patterns. Based on the result of system testing, it can be concluded that using fuzzy approach through time interval show that sequential pattern obtained is decreased because the given minimum support values are based on fuzzy time interval. The minimum support value depends on the result and time searching of fuzzy time interval sequential pattern. If sequential pattern has high support values and coverage almost all events or instance from business process then this minimum support called as the best minimum support. And the result of mining sequential pattern can be considered to change or update an existing business process.

REFERENCES

[1] R. Sarno, B. A., Sanjoyo, I. Mukhlash, H.M. Astuti, *.Journal of Theoretical & Applied Information Technology*, 54 (1), 31-38 (2013).

[2] R. Sarno, C. A. Djeni, I. Mukhlash, D. Sunaryono, *Journal of Theoretical & Applied Information Technology*, 72 (3) 412-421 (2015).

[3] R. Sarno, E.W. Pamungkas, H. Ginardi, *Proceedings of 2013 International Conference on Computer, Control, Informatics and Its Applications,* (2013), p. 319-324

[4] N. Gehrke, M. Werner., "Process Mining". An Article WISU - die Zeitschrift für den Wirtschaftsstudenten 7/13

[5] Y. Zhao, H. Zhang, S. Whu, J. Pei, L. Cao, C. Zhang and H. bohlscheid,. *ECML PKDD,* 648-663 (2009)

[6] J. Pei, J. Han, *IEEE Transactions on Knowledge and Data Engineering*, 16 (10), 881-893 (2004).

[7] Y.L. Chen, C.K. Huang. *IEEE Transactions on Systems, Man, and Cybernetics—PART B: CYBERNETICS*, 35 (5), 959-972 (2005).

[8] J. Han and M. Kamber., *Data Mining: Concept and Technique. 2nd Edition*. (Morgan-Kauffman, San Diego USA, 2006)

[9] Weske, Mathias, *Business Process Management Concepts, Languages, Architectures,* (Springer Berlin Heidelberg, New York, 2007)

[10] A. Saikhu, V. Hari, in *Proceedings of KNSI,* (2012).