Hotspot sequential pattern visualization in peatland of Sumatera and Kalimantan using shiny framework

G Abriantini[1], I S Sitanggang[1*] and R Trisminingsih[1]
[1]Computer Science Department, Bogor Agricultural University, Bogor, Indonesia

E-mail: gema.abriantini@gmail.com

Abstract. Fires on peatland frequently occurred in Sumatra and Kalimantan. Fires on peatland can be identified by hotspot sequential patterns. Sequential pattern mining is one of data mining techniques that can be used to analyse hotspot sequential patterns. Sequential pattern discovery equivalent classes (SPADE) algorithm can be applied to extract hotspot sequential patterns. The objectives of this work are: 1) to obtain hotspot sequential pattern in Sumatra and Kalimantan in 2014 and 2015, and 2) to develop a web based application using Shiny framework that is available in R package for hotspot sequential pattern visualization in peatland of Sumatra and Kalimantan. Hotspot sequential patterns were obtained using minimum support of 0.01 with the focus of analysis is the hotspot sequences with length two or more events. This work generated as many 89 sequences with length 2 or more in Sumatra in 2014, 147 sequences in Sumatra in 2015, 48 sequences in Kalimantan in 2014, and 51 sequences in Kalimantan in 2015. Hotspot sequential patterns are visualized based on peatland's characteristics, weather, and social economy. The features in this web based application have been tested and the results show that all features work properly according to the test scenario.

1. Introduction

Peatland in Indonesia has area around 20 million hectares [1]. The majority of peatland in Indonesia in located in Sumatra and Kalimantan. Ironically, peatland in Kalimantan have been burned for around 2.66 million hectares (3.58% from total). In Sumatra, around 1 million hectares' peatlands were burned. This will reduce the capability of peatland absorbing water during rainy season and this will prevent the flood occurrences.

Fires in peatland has been an annual disaster in Indonesia. Fires in peatland are usually influenced by several factors including human and weather factors. Human factors are difficult to remove because of the increasing of human needs that lead to vegetation burning for land conversion. In addition, drought is one of the weather factors that influence fires in peatlands. Fires in peatland can be detected using hotspot data. Hotspot data are collected coordinate points of area that have high temperature around those areas [1]. Hotspots that appear in sequence of at least two times in the same location are a strong indicator for peatland fires.

The emergence of hotspot that causes fires can be identified using data mining techniques. Hotspot data could be processed using sequential pattern mining which is one of the methods in data mining. There are several algorithms in sequential pattern mining, namely, generalized sequential pattern (GSP), sequential pattern discovery using equivalent classes (SPADE), frequent-pattern projected sequential pattern mining (FreeSpan), prefix projected sequential pattern mining (PrefixSpan), and mining closed frequent sequential pattern mining (CloSpan) [2].

Published under licence by IOP Publishing Ltd

Several researches in the references [3-7] have been doing researches about the processing of the hotspot dataset and those implementations. Hotspot sequence patterns were integrated with weather data in Riau to observe weather conditions for each sequence of hotspot using CloSpan algorithm [3]. There was also a research that found the patterns of hotspot sequences in Riau using PrefixSpan algorithm [4]. There was a research that detected outliers of daily hotspot data using K-Means clustering method on a web based application which was built using the Shiny framework [5]. A web based application was also developed using the Shiny framework for DBSCAN clustering of data hotspot and visualized plot of hotspot clusters in Sumatra [6]. A web based application was also built using the Shiny framework for hotspot data classification using the C5.0 algorithm [7].

The objectives of this research are to obtain hotspot sequential patterns in Peatland of Sumatra and Kalimantan and to visualize those sequential patterns in a web based application. The sequential patternswere observed based on the location of physical peatland fires. The algorithm which was applied in this research is a sequential pattern discovery using equivalent classes (SPADE). The SPADE algorithm finds patterns of sequential data quickly [8]. The SPADE algorithm already exists in R package so the sequences can be easily visualized using Shiny framework.

2. Data and methods

2.1. Dataset

This study used hotspot dataset from FIRMS MODIS NASA (http://earthdata.nasa.gov). The hotspot data in Sumatra and Kalimantan used in this study startedfrom January 2014 to December 2015. The attributes of the hotspot dataset are latitude, longitude, brightness temperature, scan, track, acq_date, acq_time, satellite, confidence, version, bright_t31, and fire radiative power. The main attributes that used in this study are latitude, longitude, acq_date, and confidence. Latitude and longitude are the attributes ofhotspots location. Acq_date is the day that those hotpots were acquired by the NASA satellite. Confidence is confidence degree of hotspot.

In addition to hotspot data, this work utilizes Sumatra and Kalimantan peatland maps in shapefile format (.shp) that were obtained from Wetland International. Those maps are used to select hotspots that are located only in peatland, and to determine the physical characteristics of peatland where hotspot regularly occurred. QuantumGIS was used in this study to select hotspots that are only located in peatland. Characteristics of the peatland in Sumatra and Kalimantan were obtained from the peatland layer in shapefile format. There are three physical characteristics of the peatland namelypeatlanddepth, peatland type, and landuse.

2.2. Sequential pattern mining

Hotspot data will continue to grow in size over time. A database that consists of a group of sequences of events recorded based on the valid time is called a time-series database [8]. Hotspot dataset which have time attribute is an example of time-series databases. There are four patterns that can be obtained by processing the time-series data, i.e. trend analysis, similarity search, sequential pattern, and periodical pattern [9].

Sequential pattern is a sequence of consecutive item-set appearance and all of the items occurred in almost the same time [8]. A sequence ‹a_1, a_2, …, a_n› is also included into the sequence ‹b_1, b_2, …, b_m›, if i_1< i_2< … < i_n is on ‹a_1 b_{i1}, a_2 b_{i2}, …, a_n b_{in}›. For example, the sequence ‹(3)(6, 7, 9)(7, 9)› can be called as part of the sequence ‹(2)(3)(6, 7, 8, 9)(7, 9)› in which (3) ⊆ (3), (6, 7, 9) ⊆ (6, 7, 8, 9), (7, 9) ⊆ (7, 9). However, ‹(2) (3)› is not part of the ‹(2, 3)› because ‹(2) (3)› means that 3 occurs after 2 and ‹(2, 3)› means 2 coincides with 3. A sequence is called to be maximal sequence if it is not included in any sequence [9]. As an illustration in the hotspot dataset, ‹(2) (3)› indicates the occurrence of hotspots in the time period 2 followed by the occurrence in the time period 3. Moreover, the sequence ‹(2, 3)› denotes occurrences of hotspot in the time period 2 and 3.

Sequential pattern mining is a process of extracting sequential pattern that its support value exceeds the specified minimum support [2]. Sequential data typically include large dataset, so it would be more efficient if pruning their sequences are performed by support values. Minimum support value is specified by user according to the needs of each user. Support is the percentage of transaction that contains a certain sub-sequences of total transactions.

2.3. SPADE algorithm

Sequential Pattern Discovery Algorithm using Equivalence classes (SPADE) is an algorithm to find patterns that use sequential equivalence classes to be divided the main problem into sub-problems that can be solved separately using join operations [8]. The input of SPADE algorithm is a sequential dataset. An example of sequential data is given in table 1.

Table 1. Example of sequential data.

Sequence ID	Event ID	Size	Items
1	10	2	{C,D}
1	15	3	{A,B,C}
1	20	3	{A,B,F}
1	25	4	{A,C,D,F}
2	15	3	{A,B,F}
2	20	1	{E}
3	10	3	{A,B,F}
4	10	3	{D,G,H}
4	20	2	{B,F}
4	25	3	{A,G,H}

Steps of the SPADE algorithm are as follows [8]:

1. Calculate the support value of all the items.
2. Calculate all items up to a maximum length sequences until no more sequences exceeding the specified minimum support.
3. Conduct the decomposition of a class of all sequences based on the length of each sequences.
4. Enumerate the entire sequence to generate a new sequence.

2.4. Research steps

Stages of this research are data pre-processing, generating hotspot sequential pattern, selecting hotspot sequential pattern, visualizing hotspot sequential pattern, and testing the application. Data pre-processing consists of several steps including data selection, data transformation, and sequential data generation. Data selection was performed to select the attributes used for mining process and to select hotspots in the study area namely peatland of Sumatra and Kalimantan. Data transformation was done to prepare sequential data of hotpots whereas sequential data generation was conducted to prepare sequential data of hotspot that its format meets the SPADE algorithm that is available on the package 'arulesSequence' in R software.

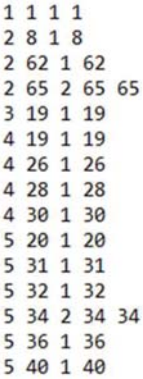
2.5. Data pre-processing

In this research, there are 3 stages on data pre-processing:

- Data selection, this stage was done to select all hotspots that occurred in peatland of Sumatra and Kalimantan. Longitude and latitude were rounded to two decimal digits in order to get several hotspots in the area with radius about 1 km. In addition, hotspots were selected only those with confidence value greater than or equal to 70% because according to the forest expert, hotspots at this confidence level have high possibility to be real fires.
- Data transformation, in this stages we transform the dataset into data format properly in the next stage. In this step, we create new attribute size and sequence ID (SID). In hotspot

dataset, size specifies the number of hotspots that occur at the same location and date. SID attribute represents the code for each different location.

- Creating sequential data, we prepare input data for the SPADE algorithm to generate the sequential pattern as shown in figure 1. In the figure 1, the columns from left to right are the SID, date_code, size, and items. SID is sequence ID. Column items will appear as many the value in the column size. Date_code represents code of the date when the hotspot occurred. The code is represented in integer. The column items represent one or more date_code indicating the sequences of hotspot occurrences.



```
1 1 1 1
2 8 1 8
2 62 1 62
2 65 2 65 65
3 19 1 19
4 19 1 19
4 26 1 26
4 28 1 28
4 30 1 30
5 20 1 20
5 31 1 31
5 32 1 32
5 34 2 34 34
5 36 1 36
5 40 1 40
```

Figure 1. Sequential dataset of hotspots, from left to right are SID, date_code, size, and items

2.6. Hotspot sequential pattern generation

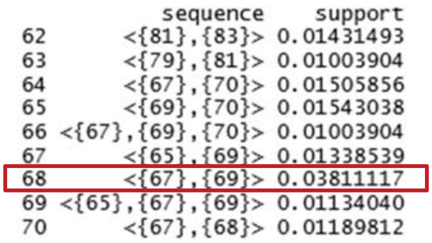
Hotspot data as the results of pre-processing stage were processed using R by applying the SPADE algorithm. In this research, the minimum support value of 0.01 was used to obtain sequential patterns at least two hotspot occurrences. Number of sequential patterns that were generated using the SPADE algorithm is provided in table 2.

Table 2. Number of sequential pattern.

Dataset	1-Sequence	2-Sequence	3-Sequence	4-Sequence
Sumatra 2014	61	24	4	-
Sumatra 2015	62	70	14	1
Kalimantan 2014	41	7	-	-
Kalimantan 2015	44	7	-	-

2.7. Hotspot sequential pattern selection

Our analysis focuses on the sequential patterns that have length of 2 or more events. This is based on consideration that hotspots which occur sequentially in at least two days are considered as strong indicator for burning peatland. Some examples of 2-sequences of hotspot sequential pattern are shown in figure 2.



```
              sequence    support
62         <{81},{83}> 0.01431493
63         <{79},{81}> 0.01003904
64         <{67},{70}> 0.01505856
65         <{69},{70}> 0.01543038
66 <{67},{69},{70}> 0.01003904
67         <{65},{69}> 0.01338539
68         <{67},{69}> 0.03811117
69 <{65},{67},{69}> 0.01134040
70         <{67},{68}> 0.01189812
```

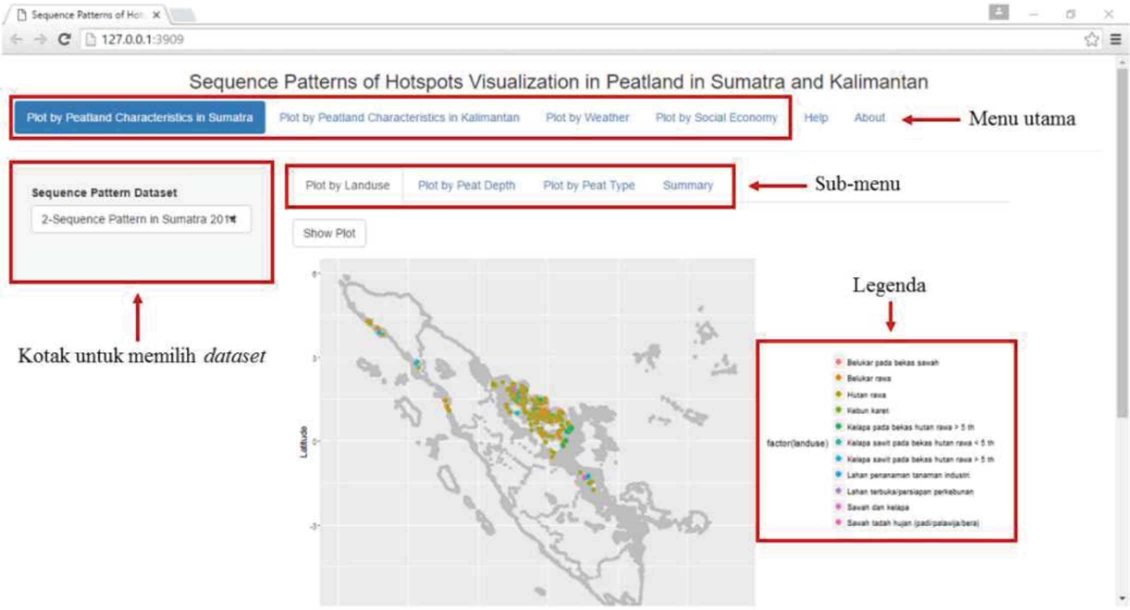
Figure 2. Hotspot sequential pattern.

In Fig. 2, the 2-sequence <{67},{69}> has the highest support value of 0.03811117. The date code 67 means 11 March 2014, whereas the date code 69 means 13 March 2014. The 2-sequence <{67},{69}> states that several hotspots were occurred on 11 March 2014 and then the occurrences were followed by other hotspots on 13 March 2014. The pattern is supported by about 3.81117 % sequence data of hotspot.

2.8. Hotspot sequential pattern visualization

A web based application for visualization of hotspot sequence patterns was built using Shiny framework. The application consists of two main files, namely Server.R as a server file and UI.R as an interface of the application. Data input of this application is the dataset in csv format that includes longitude and latitude of hotspot in the sequence patterns in peatland of Sumatra and Kalimantan.

There are four main features in the application, 1) plotting sequence patterns based on peatland's characteristics in Sumatra, 2) plotting sequence patterns based on peatland's characteristics in Kalimantan, 3) plotting sequence pattern based on weather data, and 4) plotting sequence pattern based on socio-economic data. There are additional features in the application that display a summary table for easily interpretation of sequence patterns. As shown in figure 3, the application has several features including selecting datasets, main menu, sub-menu, and legend of each plot. Each colour in peatland map represents the peatland's characteristics, such as type of peatland, peatland depth, and landuse type where the hotspot sequence occurred. Figure 4 shows a plot of hotspot sequences based on the peatland depth. Another feature in the application is summary table of the dataset that is provided in figure 5.



Sequence Patterns of Hotspots Visualization in Peatland in Sumatra and Kalimantan

Sequence Pattern Dataset: 2-Sequence Pattern in Sumatra 201

Sequence Patterns of Hot ×
127.0.0.1:3909
Plot by Peatland Characteristics in Sumatra
Plot by Peatland Characteristics in Kalimantan
Plot by Weather
Plot by Social Economy
Help
About
Menu utama
Plot by Landuse
Plot by Peat Depth
Plot by Peat Type
Summary
Sub-menu
Show Plot
Kotak untuk memilih *dataset*
Legenda
Latitude
factor(landuse)
Belukar pada bekas sawah
Belukar rawa
Hutan rawa
Kebun karet
Kelapa pada bekas hutan rawa > 5 th
Kelapa sawit pada bekas hutan rawa < 5 th
Kelapa sawit pada bekas hutan rawa > 5 th
Lahan penanaman tanaman industri
Lahan terbuka/persiapan perkebunan
Sawah dan kelapa
Sawah tadah hujan (padi/palawija/bera)

Figure 3. Plotting sequence patterns based on peatland types.

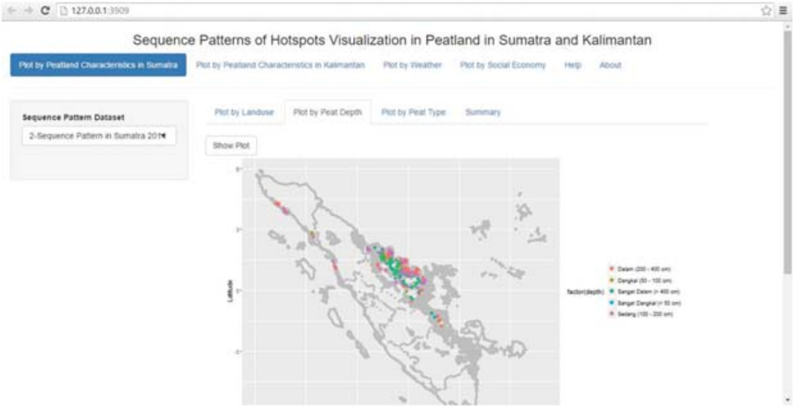
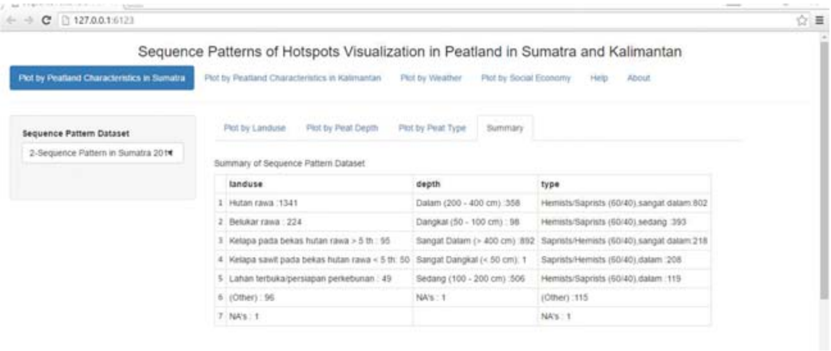


Figure 4. Plotting sequence patterns based on peatland depth.



Sequence Patterns of Hotspots Visualization in Peatland in Sumatra and Kalimantan

Summary of Sequence Pattern Dataset

	landuse	depth	type
1	Hutan rawa :1341	Dalam (200 - 400 cm) :358	Hemists/Saprists (60/40),sangat dalam:802
2	Belukar rawa : 224	Dangkal (50 - 100 cm) : 98	Hemists/Saprists (60/40),sedang :393
3	Kelapa pada bekas hutan rawa > 5 th : 95	Sangat Dalam (> 400 cm) :892	Saprists/Hemists (60/40),sangat dalam:218
4	Kelapa sawit pada bekas hutan rawa < 5 th: 50	Sangat Dangkal (< 50 cm): 1	Saprists/Hemists (60/40),dalam :208
5	Lahan terbuka/persiapan perkebunan : 49	Sedang (100 - 200 cm) :506	Hemists/Saprists (60/40),dalam :119
6	(Other) : 96	NA's : 1	(Other) :115
7	NA's : 1		NA's : 1

Figure 5. Summary of the dataset.

2.9. Application testing

The testing phase was carried out with several test scenarios and compared results with dataset visualization in Quantum GIS. The main features and additional features are worked properly and its outputs met the test scenario. The testing results are shown in table 3.

Table 3. Test scenarios.

Features	Test Scenarios	Expected Output	Work Properly?
Plotting Peatland Characteristics in Sumatra	Users select the "Show Plot"menu	Peatland map of Sumatra with a plot based on landuse, the type of peat, and peat depth is appeared	Yes
Plotting Peatland Characteristics Kalimantan	Users select the "Show Plot" menu	Peatland map of Kalimantan with a plot based on landuse, the type of peat, and peat depth is appeared	Yes
Plotting sequences by Weather	Users select the "Show Plot"menu	Peatland map of Sumatra with a plot based on average temperature, average humidity, and precipitation is appeared	Yes
Plotting sequences by Social Economy	Users select the "Show Plot"menu	Peatland map of Sumatra with a plot based on fire on land, income source, and number of school is appeared	Yes
Summary	Users select the Summary tab	Summary table of each dataset is displayed	Yes

3. Conclusion

This work results a web-based application for hotspot sequences visualization using the framework Shiny. Main features in the application includes plotting sequence patterns of hotspots in form of map based on weather, socio-economics and the characteristics of the peatland in Sumatra and Kalimantan including peat types and peat depth. Additional feature of the application is the summary table for displaying the brief description of each dataset. All features have been tested and the testing results based on the test scenarios shows that those features can work properly. This application has advantages in providing the summary of hotspot sequences and visualization of hotspot distribution in sequences. The information is important for identifying hotspots that are considered as strong indicators of peatland fires. A challenge for further study is how to improve the application by developing an additional feature that can automatically generate sequences from hotspot datasets.

Acknowledgments

The authors would like to thank to Directorate of Research and Community Service, Ministry of Research, Technology and Higher Education, Republic of Indonesia for the research grant (Contract No. 601/IT3.11/PN/2016).

References

[1] [KLH] Kementrian Lingkungan Hidup 2006 Koordinasi kelembagaan pengelolaan lahan gambut di Indonesia (Online) Available: http://www.menlh.go.id/koordinasi-kelembagaan-pengelolaan-lahan-gambut-di-indonesia

[2] Han J, Pei J and Yan X 2005 Sequential pattern mining by pattern-growth: principles and extension*s J. StudFuzz* **180** 183–220

[3] Agustina T and Sitanggang I S 2015 Sequential patterns for hotspot occurences based weather data using Clospan algorithm*Proc.3rdInternational Conference on Adaptive and Intelligent Agroindustry* (Indonesia: Bogor) p 301–305

[4] Nurulhaq N Z and Sitanggang I S 2015 Sequential Pattern Mining on hotspot data in Riau province using the Prefix pan algorithm *Proc.3rdInternational Conference on Adaptive and Intelligent Agroindustry* (Indonesia: Bogor) p 257–260

[5] Suci A M Y A and Sitanggang I S 2016 Web-based application for outliers detection on hotspot data using K-Means clustering algorithm and Shiny framework *Conf. IOP Conference Series: Earth and Environmental Science 31*

[6] Hermawati R and Sitanggang I S 2016 Web-based clustering application using Shiny framework and DBSCAN algorithm for hotspots data in peatland in Sumatra *Proc.Procedia Environmental Sciences 33* (Indonesia) p 317–323

[7] Siknun G P and Sitanggang I S 2016 Web-based classification application for forest fire data using the shiny framework and the C5.0 algorithm *Proc.Procedia Environmental Sciences 33* (Indonesia) p 332–339

[8] Zaki M J 2001 SPADE: An efficient algorithm for mining frequent sequences *J. Machine Learning* **42** p 31–60

[9] Zhao Q and Bhowmick S S 2003 Sequential pattern mining: A survey *J. Technical Report CAIS* 118