Research on $PM_{2.5}$ time series characteristics based on data mining technology

Lifang Zhao[1] and Jin Jia

Department of Environmental Science, Zhejiang University, Hangzhou 310058, China

[1] alanskying@163.com

Abstract. With the development of data mining technology and the establishment of environmental air quality database, it is necessary to discover the potential correlations and rules by digging the massive environmental air quality information and analyzing the air pollution process. In this paper, we have presented a sequential pattern mining method based on the air quality data and pattern association technology to analyze the $PM_{2.5}$ time series characteristics. Utilizing the real-time monitoring data of urban air quality in China, the time series rule and variation properties of $PM_{2.5}$ under different pollution levels are extracted and analyzed. The analysis results show that the time sequence features of the $PM_{2.5}$ concentration is directly affected by the alteration of the pollution degree. The longest time that $PM_{2.5}$ remained stable is about 24 hours. As the pollution degree gets severer, the instability time and step ascending time gradually changes from 12-24 hours to 3 hours. The presented method is helpful for the controlling and forecasting of the air quality while saving the measuring costs, which is of great significance for the government regulation and public prevention of the air pollution.

1. Introduction

In recent years, the problem of atmospheric environmental pollution is becoming more and more serious with the rapid development of economy in China. The annual average of $PM_{2.5}$ concentration in some areas is much higher than the WHO annual standard, especially in the central and eastern coastal regions. Among the various pollutants, $PM_{2.5}$ is one of the core pollutants during the atmospheric complex pollution process[1,2]. Generally speaking, $PM_{2.5}$ is a mixture of particulate matter and secondary particles with complex composition. A variety of gas, solid precursor SO_2, NO_x, elemental carbon (EC), organic carbon (OC) can be converted into $PM_{2.5}$ through the homogeneous and heterogeneous reaction. The atmospheric compound pollution not only threatens the national health and social stability, but also restricts the economic development of many industrial provinces.

Based on the air quality data and statistical methods, many researches about air pollution process have got carried out in decades, including regional division, spatial and temporal prediction model, transformation relationship between the pollutants and precursors, human health assessment and the environmental management technology. In Japan, researchers [3] examined the SPM and $PM_{2.5}$ levels per hour for the 1990 monitoring stations in Tokyo, and analyzed the positive correlation relationship between vehicle traffic and $PM_{2.5}$. The European researchers [4] utilized the coarse grain data from traffic and background monitoring stations in three major cities including London, Madrid and Athens, in order to analyze the seasonal, weekly and diurnal $PM_{2.5}$ variation properties, and the correlations with traffic and wind speed. In UK, researchers[5] used the $PM_{2.5}$ hourly monitoring data to figure out

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

the correlations among the average circadian $PM_{2.5}$ distribution, seasonal variations and meteorological conditions. As for the analyses of $PM_{2.5}$ pollution characteristics, the domestic and foreign scholars mainly focused on the study of the pollution mechanism and principle, such as the microscopic analyses of the $PM_{2.5}$ components and formation mechanism [6]. Currently, the main theoretical methods used for the air quality data analyses include traditional statistics computation, forecasting classification based on data mining, clustering and correlation analysis technology [7,8].

With the development of environmental engineering technology, the automatic monitoring and recording system has been widely used in the field of air pollution prevention and control, which has accumulated a large amount of atmospheric environmental data. Utilizing the correlation analysis technology to dig and analyze the massive environmental air quality information offers an effective solution for the illumination of the mechanism and principle during the atmospheric pollution process. $PM_{2.5}$ is one of the key atmospheric pollutants, the research on the $PM_{2.5}$ sequential properties will certainly promote the analyses and control of the air pollution problems. Different from the traditional analysis method of air quality, we have presented a sequential pattern mining method based on the air quality data and pattern association technology to analyze the $PM_{2.5}$ time series characteristics. The presented method systematically integrates the data acquisition, format conversion, discretization processing, arules algorithm and pattern evaluation together, in order to build an air quality hourly data mining platform. The study of this paper is helpful for the controlling and forecasting of the air quality while saving the measuring costs, which is of great significance for the government regulation and public prevention of the air pollution.

2. Methods

In this paper, we combine the association rules mining method with big data theory and establish a sequential pattern discovery method based on air quality data, which utilizes the data mining correlation analysis technique to analyze the atmospheric pollution process. The presented method integrates the automatic collection of real-time data, format conversion, data discretization processing, arules series algorithm and data identification together to build an air quality hourly data mining platform.The principle graph of the presented sequential pattern discovery method can be divided into four key procedures including construction of market basket database, sequential pattern mining, sequential pattern recognition and results resolution, as is shown in Figure 1.

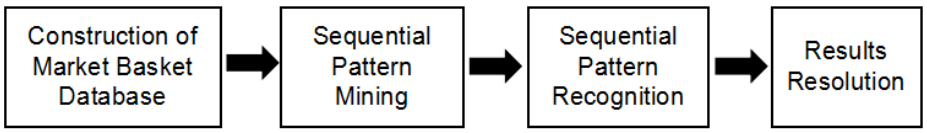


Figure 1. The principle graph of the presented sequential pattern discovery method.

2.1. Construction of the market basket database

2.1.1 Automatic data collection The data source used in this study is recorded from the formal air quality website (http://www.pm25.in/), which provides the air quality real-time data from the Ministry of Environmental Protection. In order to realize the automatic collection and storage of data, we design a data acquisition and storage software tool using the script recording method. By designing the data acquisition and storage process, the computer can automatically logs into the website per hour through the API interface and store the air quality data continuously.

2.1.2 Data pre-processing The data analysis and result output process is performed using the R software, which is a open source software and widely used in the data statistic analysis. In order to use the R tool for analyses, the collected air quality data need to be pre-processed. Firstly, the data is transformed from text format into the table database format. Then, the $PM_{2.5}$ concentration data is

extracted and discretized from the large amount of air quality data. The specific discretization scheme is based on the National Ambient Air Quality Index (HJ633-2012), as is shown in Table 1.

Table 1. Standard for discretization of $PM_{2.5}$ concentration data. The unit is $\mu g/m^{3..}$

$PM_{2.5}$ Concentration	0-35	36-75	76-115	116-150	151-250	>300
Level	1	2	3	4	5	6

2.2. Sequential pattern mining

2.2.1 The Principle and algorithm of sequential pattern mining The sequence pattern mining is a classical sequence analysis method used for the identification of dynamic systems or recurring features to predict future specific events, which not only emphasizes the relevance, but also focuses on the order of occurrence time. The support degree of a sequence *s* is the proportion of all the data sequences that contain *s*. If the support degree of s is greater than or equal to a user-specified threshold minsup, then s is declared to be a sequential pattern. Given a sequence data set D and a user-specified minimum support threshold, then the task of sequential pattern discovery is to find all sequences with support degree satisfying s≥*minsup*. In this paper, we use the Apriori algorithm to carry out the sequence pattern mining process. The basic structure of the Apriori algorithm used in this paper is shown in Table 2. The algorithm will iteratively generate new candidate sequences, subtract nonpolar candidates, and then count the remaining candidate items to identify the strong sequence pattern.

Table 2. The Apriori algorithm used for sequential pattern discovery.

1. k1
2. F_k={i | i_I } { Find all frequent 1-subsequences }
3. repeat
4. k=k+1
5. C_k= apriori-gen (F_{k-1}) { Generate candidate k-sequences }
6. for each data sequence tTdo
7. C_t=subsequence(C_k, t) { Identify all candidates contained in t.}
8. for each candidate k- sequence cC_t, do
9. { Increment the support count.}
10. end for
11. end for
12. Fk={c|c∈Ck σ(c)/N≥minsup} } { Extract the frequent k-sequences }
13. until F_k=Ø
14. answer=∪F_k

2.2.2 Mining method In order to obtain a valid sequence pattern more efficiently, time constraints can be imposed on the events and elements of the pattern. The constrains usually include maximum span constraints, minimum time intervals, maximum spacing constraints, and window size constraints. The maximum span constraint refers to the maximum time difference between the latest and the earliest occurrences of the events allowed in the entire sequence. The minimum interval (*mingap*) and the maximum time interval (*maxgap*) refers to the minimum and maximum values of the time difference between two successive elements. The window size threshold is the maximum allowable time difference between the latest and the earliest occurrences of the event in any element of the sequence pattern. After generating a large number of sequence patterns, the pattern needs to be evaluated suing the parameter of support degree. Because the support degree of each sequence is different, we need evaluate the strength of the sequence under different time intervals.

According to the monitoring data during recent years, $PM_{2.5}$ pollution occurred mainly in the autumn and winter. We study the time series characteristics of $PM_{2.5}$ based on the air quality data

collected from September 4, 2013 to December 4 using the presented sequential pattern discovery method. The total length of the collected data timestamp is more than 100 days, so the time window is designed to be 1-3168 hours, the parameters of the time window settings is shown in Table 3. The support degree is set to be 0.3 to ensure the effectiveness of the results.

Table 3. Parameters of the sequential pattern mining process.

Time Window (h)	Mingap	Maxgap	Max-maxlen	Maxlen	Maxsize
1	0	1	2	2	7
2	1	2	3	2	7
3	2	3	4	2	7
6	5	6	7	2	7
12	11	12	13	2	7
24	23	24	25	2	7
48(2d)	47	48	49	2	7
96(4d)	95	96	97	2	7
192(8d)	191	192	193	2	7
384(16d)	383	384	385	2	7
768(32d)	767	768	769	2	7
1536(64d)	1535	1536	1537	2	7
3168(132d)	3167	3168	3169	2	7

2.2.3 Programming and computing The ArlesSequences package in R software is used to calculate the data sequence. The sequential pattern mining in R data mining consists of five steps, including data inputting, data preprocessing, sequential pattern mining, pattern evaluation and results output. Figure 2 shows the specific R language implementation process for the sequential pattern mining.

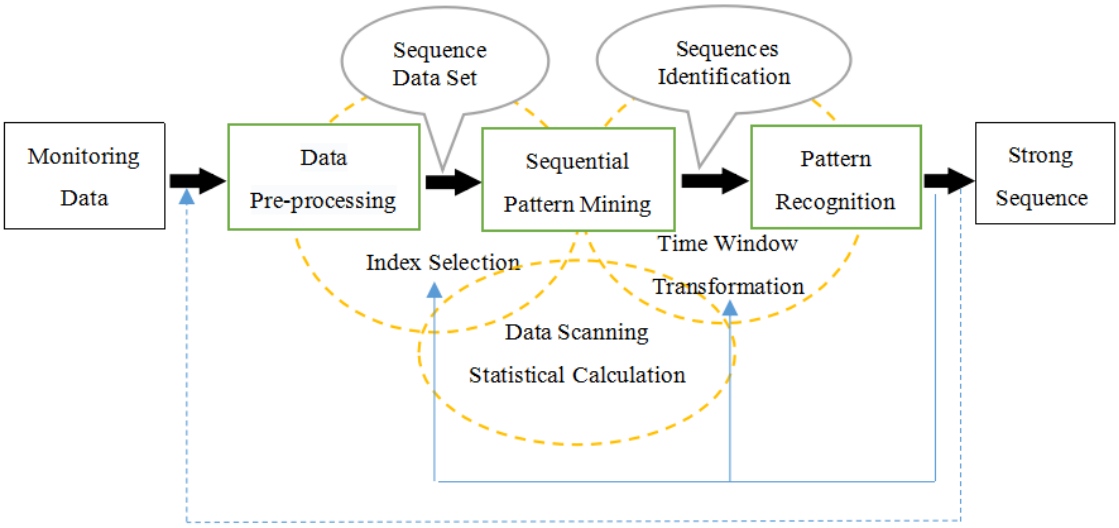


Figure 2. The implementation process of sequential pattern mining in R software.

The first and second step during the implementation process have been completed in the market basket database construction and mining method section. The third step is sequential pattern mining, including ArulesSquence package calling, data set scanning and the sequence support calculation. The fourth step is to evaluate the obtained patterns, which needs to filter and remove the superfluous

patterns according to the task requirements and performs the pattern characterization process. Finally, based on the sequences selection results, the $PM_{2.5}$ sequence feature under different time intervals during the atmospheric composite pollution process can be obtained.

3. Results and discussions

According to the mining scheme and the predetermined parameters, the PM2.5 time sequence patterns of different grades is obtained. The specific results of the sequence patterns are shown in the following tables.

Table 4. The support value under each time differences when the $PM_{2.5}$ pollution level keeps stable. "NA" indicates that the support value is smaller than 0.3.

Time Difference	{$PM_{2.5}$=3} to {$PM_{2.5}$=3}	{$PM_{2.5}$=4} to {$PM_{2.5}$=4}	{$PM_{2.5}$=5} to {$PM_{2.5}$=5}	{$PM_{2.5}$=6} to {$PM_{2.5}$=6}
1h	0.98	0.94	0.87	0.64
2h	0.95	0.89	0.80	0.54
3h	0.94	0.87	0.78	0.50
6h	0.91	0.81	0.71	0.40
12h	0.85	0.71	0.52	0.33
24h	0.47	NA	NA	NA

Table 5. The support value under each time differences when the $PM_{2.5}$ pollution grade changes from low level to high level . "NA" indicates that the support value is smaller than 0.3.

Time Difference	{$PM_{2.5}$=2} to {$PM_{2.5}$=3}	{$PM_{2.5}$=3} to {$PM_{2.5}$=4}	{$PM_{2.5}$=4} to {$PM_{2.5}$=5}	{$PM_{2.5}$=5} to {$PM_{2.5}$=6}
1h	0.93	0.84	0.79	0.49
2h	0.88	0.81	0.77	0.50
3h	0.88	0.80	0.76	0.50
6h	0.86	0.78	0.71	0.42
12h	0.82	0.66	0.65	0.38
24h	0.56	0.32	0.32	NA

Table 6. The support value under each time differences when the $PM_{2.5}$ pollution grade changes from high level to low level. "NA" indicates that the support value is smaller than 0.3.

Time Difference	{PM2.5=3} to {PM2.5=2}	{PM2.5=4} to {PM2.5=3}	{PM2.5=5} to {PM2.5=4}	{PM2.5=6} to {PM2.5=5}
1h	0.91	0.83	0.79	0.50
2h	0.86	0.80	0.77	0.48
3h	0.85	0.79	0.74	0.46
6h	0.83	0.76	0.70	0.40
12h	0.79	0.66	0.62	NA
24h	0.40	NA	NA	NA

Firstly, we focus on the variation of the support value when the $PM_{2.5}$ levels keeps stable. Table 4 shows the support value under each time differences when the $PM_{2.5}$ pollution level keeps stable. When the time difference is 1 h, the support degree of the 3rd-level $PM_{2.5}$ remaining stable after 1 is

high, which has a value of 0.98. Meanwhile, the support degree of the 6th-level $PM_{2.5}$ remaining stable is 0.64. The particles under low pollution level are likely to last for almost 1 h. With the increase of the time difference, the support degree of the 3rd-level $PM_{2.5}$ remaining stable after 12 h becomes 0.85, while the support degree of 3rd-level is 0.33. The instability time of the 3rd and 4th levels is longer than 12 h, while the instability time of the 6th level is about 3 h. When the time difference reaches 24 h, most of the support degree is smaller than 0.3, which suggests that 24 h can be defined as a typical time length for the $PM_{2.5}$ pollution to maintain stability.

Secondly, the $PM_{2.5}$ concentration need take some particular time to get changed. As is shown in Table 5, the support degree under each $PM_{2.5}$ pollution level decreases evidently when the time difference is 12 to 24 h, indicating that the general ascending time of the $PM_{2.5}$ concentration is 12-24 h. Furthermore, it can be seen from table 6 that the instability time and step ascending time gradually changes from 12-24 hours to 3 hours when the pollution degree gets severer. The typical descending time of $PM_{2.5}$concentration below 5th level is 12-24 h, and the typical descending time from 6th level to 5th level is 6-12 h. In this paper, we mainly utilize the sequential pattern mining method to analyze the time series rule and variation properties of $PM_{2.5}$, which provides an effective method for the resource utilization of the massive air quality data. Future work would update more air pollution data and optimize the mining algorithm and parameters to further improve the validity and accuracy of the presented method.

4. Conclusions

In this paper, we have presented a sequential pattern mining method based on the air quality data and pattern association technology, and analyzed the time series rule and variation properties of $PM_{2.5}$ at different pollution levels through utilizing the real-time monitoring data of urban air quality. An air quality hourly data mining platform is built by systematically integrating the data acquisition, format conversion, discretization processing, arules algorithm and pattern evaluation together. The analysis results show that the time sequence features of the $PM_{2.5}$ concentration is directly affected by the alteration of the pollution degree. This paper provides an effective method for the resource utilization of the massive air quality data, which is helpful for the controlling, forecasting and mechanism analyses of the air pollution.

References

[1] Zhang Y, Yao Z, Shen X, Liu H, and He K 2015 *England. Atmos. Environ.* **885-891** 122
[2] Shen Z, Sun J, Cao J, Zhang L, Zhang Q, and Lei Y 2016 *Netherlands. Sci. Total. Environ.* **569-570** 619
[3] Kunio Hara, Junichi Homma, Kenji Tamura, Mariko Inoue, Kanae Karita, and Eiji Yano 2013 *United States. J. Air. Waste. Manage.* **737** 63
[4] P Kassomenos, S Vardoulakis, A Chaloulakou, G Grivas, R Borge, and J Lumbreras 2012 *England. Atmos. Environ.* **337-347** 54
[5] Harrison, Roy M, Laxen, Duncan, Moorcroft, and Stephen 2012 *England. Atmos. Environ.* **115-124** 46
[6] Vienneau D, and D J Briggs 2013 *United States. Environ. Int.* **106-115** 51
[7] Zheng J, Hu M, Peng J, Wu Z, Kumar P, and Li M 2016 *England. Chemosphere.* **480-487** 159
[8] Vaidyanathan A, Dimmick W F, Kegler S R, and Qualters J R 2013 *England. Int. J. Health. Geogr.* **12** 12