Stream Data Analysis of the Weather on the Location using Principal Component Analysis

주성분 분석을 이용한 지역기반의 날씨의 스트림 데이터 분석

Kim, Sang Yeob[1] · Kim, Kwang Deuk[2] · Bae, Kyoung Ho[3] · Ryu, Keun Ho[4]
김상엽 · 김광덕 · 배경호 · 류근호

Abstract

The recent advance of sensor networks and ubiquitous techniques allow collecting and analyzing of the data which overcome the limitation imposed by time and space in real-time for making decisions. Also, analysis and prediction of collected data can support useful and necessary information to users. The collected data in sensor networks environment is the stream data which has continuous, unlimited and sequential properties. Because of the continuous, unlimited and large volume properties of stream data, managing stream data is difficult. And the stream data needs dynamic processing method because of the memory constraint and access limitation. Accordingly, we analyze correlation stream data using principal component analysis. And using result of analysis, it helps users for making decisions.

Keywords: Stream data, Sensor network, sequential properties, principal component analysis.

1. Introduction

Recently, by development in the sensor networks and ubiquitous techniques, the decision of user is reflected by the data which is collected and analyzed at real-time without constraint of time and spatial area. Also, analysis and prediction of the collected data can provide useful information to users(Manjeshwar, 2001 ; Kim, 2007).

The data collected from the sensor network is defined as the stream data which has infinite and continuous properties. For example, the weather data collected in real-time or off-line by person and measurement equipment or artificial satellite such as temperature, humidity, sunshine, and so on. The continuously occurring in data is defined as stream data.

However, one of the difficulties facing the problem of analyzing stream data is storing and managing all of these data because of the large volume.

Analyzing and processing stream data have some problems due to the properties of stream data. The stream data has the continuous and infinite properties by time, and use limited memory. So, the stream data changes dynamically, therefore continuous processing methods are needed. Moreover, the random access to the data is impossible because stream data has sequential property by time. Therefore, in order to process such data easily, we apply the sliding window technique and normalization method. And we analyze the stream data using principal component analysis.

2. Related Work

The stream data has continuous property and it almost makes the unlimited characteristic occurs in data. Also it has

[1] Researcher New & Renewable Energy Research Group, Korea Institute of Energy Research (E-mail:sykim@kier.re.kr)
[2] Group Manager · New & Renewable Energy Research Group · Korea Institute of Energy Research (E-mail:kdkim@kier.re.kr)
[3] Researcher · Research Institute of Geoinformatics, Korea Association of Surveying & Mapping (E-mail:qpandora@paran.com)
[4] Corresponding author · professor, Division of Computer Engineering, Chungbuk National University (E-mail:khryu@dblab.cbnu.ac.kr)

the feature of keeping its order at real-time. In addition it is difficult to store the whole stream of data in database; therefore, it observes the data from memory at real-time in order to satisfy the need for continuous queries from users(Lee, 2007). Fig. 1 shows the characteristic of stream data. The time-series data is recoded data according to time with the interval in about one event or various events. Stream data can also be seen as time-series data for a schedule of time, it is possible to classify the data which is collected in intervals.

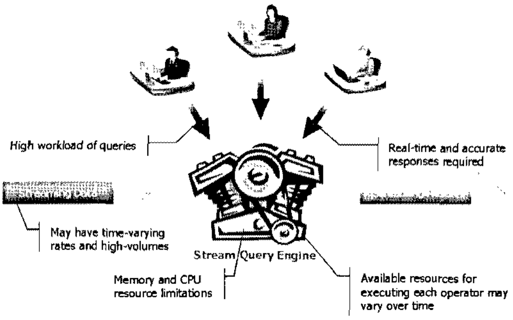

High workload of queries
Real-time and accurate responses required
May have time-varying rates and high-volumes
Stream/Query Engine
Memory and CPU resource limitations
Available resources for executing each operator may vary over time

Fig. 1. characteristic of stream data

As an example, parting of the stock price index which changes every day, the sale volume of specific commodity by month or days, and the production volume of crops by years. The time-series data which is observed with one point of view that depends on the previous data and it analyzes observed data for discovering a model which is able to predict the values of observed data in the future.

Also, the new and renewable energy resources recently become an issue that has the characteristic of stream data as well. The new and renewable energy resources is divided into two parts, new energy and a renewable energy, the new energy defined a fuel cell, a hydrogen energy and a coal liquefaction gas. And the renewable energy defined as solar, solar photovoltaic power and biomass energy, the wind power, small hydraulic power, ocean energy, the waste energy and the geothermal energy(Ryu, 2009). The solar energy provides information which is related with a solar energy in a location with the point by establishment, in order to analyze the measured energy which is divided into essential measuring element and additional measuring element. Wind power energy achieved the supply and the technical development from the 90' s; the housing development of wind is appeared by invested private organization of commercial objective.

There are many projects for processing of new and renewable energy resources. BSRN (baseline surface radiation Network) is Radiometric Network for supporting World Climate Research Program and the other project of different scientific program. The objective of BSRN provides the Short and Longwave Surface Radatio Fluxes using a high sampling ratio. FSEC (florida solar energy center) as the solar energy research center of American Florida collects the meteorological data for the research of meteorological data, the data is collected with automatic and stored in web site. NREL (national renewable energy laboratory) is a national remaking energy laboratory of the United States, it investigates the renewable energy and researches, and develops a new energy technique.

The objective of this paper is to provide useful information to users by correlation analysis of stream data. Also it helps organizing and managing stream data and to select the optimal location related to energy facility using information valuable to user.

3. Analyzing Stream Data using Principal Component Analysis

In this section, we describe a principal component analysis (PCA) for analyzing stream data. Fig. 2 shows processing of stream data.

Before the analysis, we use a normalization method for reducing the gap of each variable.

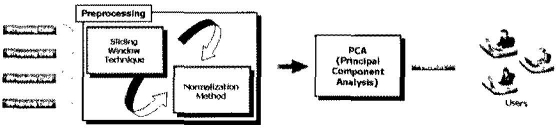


Fig. 2 processing of stream data

The normalization process calculates the average and a standard deviation of each data; it then accomplishes the normalization process in order to make the efficient analysis possible. The value of normalization N is given by the equation in formula.

$$N_i = \frac{X_i - \mu}{\sigma}$$

And we apply a sliding window technique for reducing data set' s dimension. The sliding window technique calculates the average of the data which is current point of time with maintaining window size. It compares the mean value from the mean value in the next point of time, and if the value exceeds a standard value, then uses it. In other words, it decides a standard value according to the using data and if the condition is satisfied, then the value makes to be used or not. When we use the sliding window technique, it is possible to manage the efficient storage. Also, it reduces the processing cost and storage space. Moreover, it overcomes the overlapping of data. Fig. 3 shows the sliding window technique with size 5.

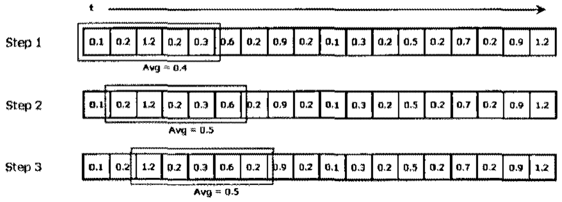
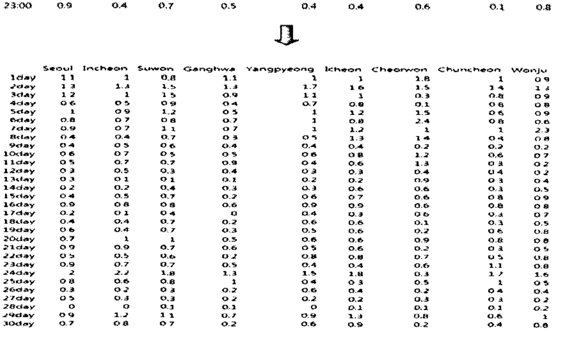


Fig. 3. example of sliding window technique

As shown in fig. 3, the stream data inputs some point of time as t continuously, it calculates the average of data with moving window according to time. The data used in this paper is the weather data set of Korea. In order to preprocess, we use the sliding window technique. The dataset which is divided by locations is shown in fig. 4. The data shows the rainfall of each location at a unit of time. Using sliding window technique, the dataset can be used in analysis with reduced data at unit of day.



	Seoul	Incheon	Suwon	Ganghwa	Yangpyeong	Icheon	Cheorwon	Chuncheon	Wonju
1:00	0.9	0.1	0.8	0.9	0.9	0.4	0.8	0.1	0.9
2:00	0.4	0.2	0.7	0.8	0.7	0.8	0.8	0.4	0.3
3:00	0.8	0.4	0.4	0.9	0.6	0.5	0.3	0.8	0.9
4:00	0.9	0.5	0.9	0.4	0.7	0.8	0.1	0.8	0.8
5:00	0.8	0.5	0.7	0.5	0.5	0.5	0.5	0.6	0.9
6:00	0.8	0.7	0.8	0.7	0.4	0.8	0.4	0.8	0.6
7:00	0.9	0.2	0.5	0.7	0.6	0.6	0.4	0.1	0.3
8:00	0.9	0.4	0.7	0.3	0.5	0.5	0.4	0.4	0.8
9:00	0.7	0.5	0.6	0.4	0.4	0.4	0.2	0.2	0.2
10:00	0.8	0.3	0.5	0.5	0.6	0.8	0.2	0.6	0.7
11:00	0.9	0.7	0.7	0.4	0.4	0.6	0.3	0.3	0.2
12:00	0.8	0.5	0.3	0.4	0.3	0.3	0.4	0.4	0.2
13:00	0.7	0.1	0.1	0.1	0.2	0.2	0.9	0.3	0.4
14:00	0.5	0.2	0.4	0.3	0.3	0.6	0.6	0.3	0.5
15:00	0.9	0.2	0.7	0.2	0.6	0.7	0.6	0.8	0.9
16:00	0.9	0.4	0.8	0.6	0.9	0.9	0.6	0.8	0.8
17:00	0.7	0.2	0.4	0.6	0.4	0.3	0.6	0.3	0.7
18:00	0.8	0.4	0.7	0.2	0.6	0.6	0.1	0.3	0.5
19:00	0.6	0.4	0.7	0.3	0.5	0.6	0.2	0.6	0.8
20:00	0.7	0.6	0.7	0.5	0.6	0.6	0.9	0.8	0.8
21:00	0.9	0.5	0.7	0.6	0.5	0.6	0.2	0.3	0.5
22:00	0.5	0.5	0.6	0.2	0.8	0.8	0.7	0.5	0.8
23:00	0.9	0.4	0.7	0.5	0.4	0.4	0.6	0.1	0.8

	Seoul	Incheon	Suwon	Ganghwa	Yangpyeong	Icheon	Cheorwon	Chuncheon	Wonju
1day	1.1	1	0.8	1.1	1	1	1.8	1	0.9
2day	1.3	1.3	1.5	1.3	1.7	1.6	1.5	1.4	1.3
3day	1.2	1	1.5	0.9	1.1	1	0.3	0.8	0.9
4day	0.6	0.5	0.9	0.4	0.7	0.8	0.1	0.8	0.8
5day	1	0.9	1.2	0.5	1	1.2	1.5	0.6	0.9
6day	0.8	0.7	0.8	0.7	1	0.8	2.4	0.8	0.6
7day	0.9	0.7	1.1	0.7	1	1.2	1	1	2.3
8day	0.4	0.4	0.7	0.3	0.5	1.3	1.4	0.4	0.8
9day	0.4	0.5	0.6	0.4	0.4	0.4	0.2	0.2	0.2
10day	0.6	0.7	0.5	0.5	0.6	0.8	1.2	0.6	0.7
11day	0.5	0.7	0.7	0.9	0.4	0.6	1.3	0.3	0.2
12day	0.3	0.5	0.3	0.4	0.3	0.3	0.4	0.4	0.2
13day	0.3	0.1	0.1	0.1	0.2	0.2	0.9	0.3	0.4
14day	0.2	0.2	0.4	0.3	0.3	0.6	0.6	0.3	0.5
15day	0.4	0.5	0.7	0.2	0.6	0.7	0.6	0.8	0.9
16day	0.9	0.8	0.8	0.6	0.9	0.9	0.6	0.8	0.8
17day	0.2	0.1	0.4	0	0.4	0.3	0.6	0.3	0.7
18day	0.4	0.4	0.7	0.2	0.6	0.6	0.1	0.3	0.5
19day	0.6	0.4	0.7	0.3	0.5	0.6	0.2	0.6	0.8
20day	0.7	1	1	0.5	0.6	0.6	0.9	0.8	0.8
21day	0.9	0.9	0.7	0.6	0.5	0.6	0.2	0.3	0.5
22day	0.5	0.5	0.6	0.2	0.8	0.8	0.7	0.5	0.8
23day	0.9	0.7	0.7	0.5	0.4	0.4	0.6	1.1	0.8
24day	2	2.2	1.8	1.3	1.5	1.8	0.3	1.7	1.6
25day	0.8	0.6	0.8	1	0.4	0.3	0.5	1	0.5
26day	0.3	0.2	0.3	0.2	0.6	0.4	0.2	0.4	0.4
27day	0.5	0.3	0.3	0.2	0.2	0.2	0.3	0.3	0.2
28day	0	0	0.1	0.1	0	0.1	0.1	0.1	0.2
29day	0.9	1.2	1.1	0.7	0.9	1.3	0.8	0.6	1
30day	0.7	0.8	0.7	0.2	0.6	0.9	0.2	0.4	0.8

Fig. 4. the data set of rainfall

The PCA is one of the techniques which analyze stream data(Kim, 2009). It has a property of liner transformation that enables it to map the data into new spaces and chooses the subspaces that have the biggest dispersion. If the unit of factor is same for the PCA, that use a covariance matrix. In the opposite case, that uses a correlation matrix. Here, the covariance matrix is the value of the higher correlation degree of 2 probability variables. If one of the values is increased, and another value is also increased, then the value of a covariance matrix becomes positive number. In the opposite case, if one of the values is increased, and another value is decreased, then the value of a covariance matrix becomes a negative number.

With the above method, we are able to analyze principal component of stream data. The data which is used in this paper is the weather data set of Korea. The data construct many factors at each location that are temperature, humidity, sunshine, rainfall, and so on. Therefore, we find the related factor by PCA and the factor have strongest relationship. Through the PCA, we find the important factor at each locations and the correlation with each factors and help users to

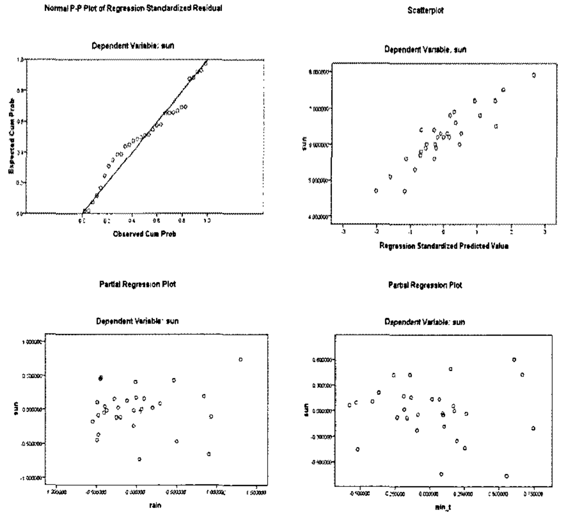


Fig. 5 example of correlation analysis

decide for optimum location of energy facility.

The correlation analysis is the degrees of relation, such as the method analyze relationship between each variable(Kim, 2009). Fig. 5 shows the example of correlation analysis. If the distribution of points becomes closely, then the correlation becomes larger. In other word, the distribution becomes scattered.

Accordingly, we manage and analyze the stream data efficiently with principal component analysis. And we find the correlations of each factor, and using useful information, it helps users for making decisions.

4. Evaluation

The embodiment environment of the method used CPU 2.66GHz Intel Core TM 2 Duo and 2.00GB RAMs. The platform used in Windows XP Professional and used SPSS 1.6. The data set is the weather data set of Korea from 1970 until 1999 and divided into two parts that are constructed locations and properties.

In order to predict and conduct correlation analysis, we accomplish two kind of experiment. First, the weather factor data set by locations are analyzed by principal component analysis. Second, the data set of each location are analyzed by multiple regression analysis.

Based on accomplished preprocessing, we analyze the stream data. First, the principal component analysis uses a covariance matrix and accomplishes the analysis about data set of each composed locations. We extract the result of analysis which is influenced location with rainfall, humidity, and sunshine. The results of PCA help users for making decisions and finding the optimal locations of buildings related to energy facilities.

And next, we analyze prediction and correlation analysis by multiple regression analysis. When the new data and influenced factors at locations are inputted, the results of analysis help the decision of user with predicting future. Also, we analyze the correlation between each factor. In addition, the method is able to efficiently manage and store stream data.

We analyze the influenced locations by rainfall, humidity, sunshine and wind velocity using PCA. Fig. 6 shows the results of rainfall.

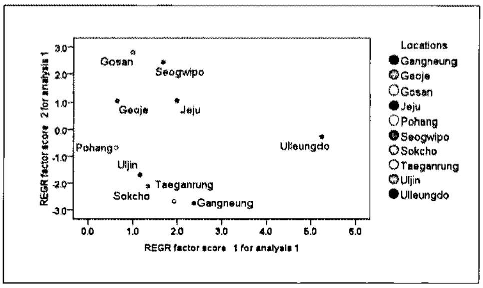


Fig. 6. results of rainfall

Fig. 7 shows the results of humidity.

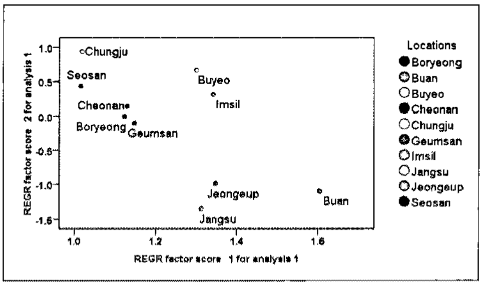


Fig. 7. results of humidity

Fig. 8 and 9 shows the sunshine and wind velocity.

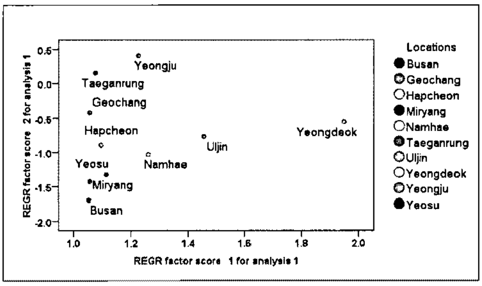


Fig. 8. results of sunshine

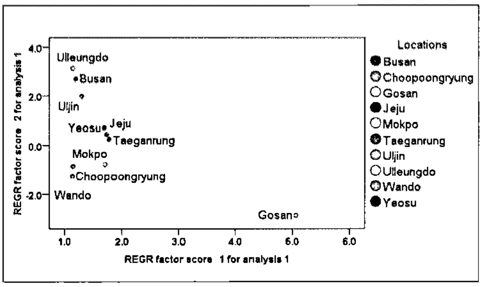


Fig. 9. results of wind velocity

Each graph is a result of analyzing x and y axis as 1^{st}, 2^{nd} principal components respectively. The cities that have high value about rainfall value such as Ullengdo, Jeju, and Gangneung. Also, the locations of our country cities have high value about humidity value such as Busan, Gosan and Jangsu and cities that have high sunshine value such as Yeongdeok, Buyeo, Busan and Namhae. Therefore, it assists in selecting the optimal location related to energy facility using information valuable to user.

5. Conclusions

Recently, the data is collected in real-time, without constraint of time and space from sensor networks and ubiquitous technologies, and is analyzed to help users for making decisions. Also, the analysis and correlation analysis of stream data provides useful information to users.

The data collected from sensor networks, is stream data that has continuous, unlimited properties. The continuously occurring in data is defined as stream data, and it can also be defined as a time-series data because of its sequential property.

In this paper, we analyze the stream data using principal component analysis and correlation analysis of stream data efficiently. We applied the sliding window technique and normalization method for preprocessing the stream data. The preprocessing overcomes the limitations in efficient managing and deleting the overlapping. The sliding window technique extracts the data for constructing regression model by calculating the difference between average of current data and past data with fixed window size and standard value.

From our experiments we find that the locations of our country's cities as Busan, Gosan and Jangsu have high humidity value and the cities such as Yeongdeok, Buyeo, Busan and Namhae have high sunshine value.

Consequently, we manage to analyze the stream data efficiently and find the correlations of each factor. And using useful information, it helps in selecting the optimum location for facilities such as the wind power plant, the solar power plant, and geothermal power plant.

Acknowledgement

This research was supported by the Korea Institute of Energy Research (KIER) and by a grant (07KLSGC02) from Cutting-edge Urban Development - Korean Land Spatialization Research Project funded by Ministry of Land, Transport and Maritime Affairs.

References

Kim, S. H. (2007), *Stream Data Prediction and Future Classification using Incremental Model Update Technique,* Chungbuk National University master's thesis.

Kim, S. Y., Kim H., Kim, K. D. and Ryu, K. H. (2009), Multi Regression Analysis for Time-series Data in Stream Environment, *IWAC 2009,* pp. 654-66.

Lee, Y. K., Jung, Y. J. and Ryu, K. H. (2007), Design and Implementation of a System for Environmental Monitoring Sensor Network, *In Proceedings of the Conference on APWeb/WAIM Workshop on DBMAN,* pp. 223-228.

Manjeshwar, A. and Agrawal, D. P. (2001), TEEN: A routing protocol for enhanced efficiency in wireless sensor networks, *International Workshop Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing,* pp. 2009-2015.

Ryu, K. H., Lee, S. H., Kim, H. S., Kim, S. Y., Park, S. K. and Jo, Y. B. (2009), *The Technology of Real Time Monitering and Processing for managing New and Renewable Energy Resource,* Research Report, Korea Institute of Energy Research.

(접수일 2010. 02. 24, 심사일 2010. 02. 24, 심사완료일 2010. 04. 02)