Deep Data Anaylizing Application Based on Scale Space Theory in Big Data Environment

Ye Hu
College of Mechanical Engineering
Ningbo Dahongying University
Ningbo, China
Tongji University, Shanghai, China
pxhuye@126.com

Kun Gao
Zhejiang Business Technology Institute, China
Zhejiang Wanli University
Ningbo, China
kungao@live.com

Zheng Xu
Tsinghua University, Beijing, China
The Third Research Institute of the Ministry of Public Security,
Shanghai,China
xuzheng@shu.edu.cn

***Abstract*—This paper introduces the basic scientific idea of the multi-scale to the field of big data analyzes, proposes a multi-scale framework of data analyzes in big data environment, present the multi-scale algorithm framework of knowledge conversion theory and apply the algorithm framework to the multi dimension association rules analysis. The proposed multi-scale association rule analysis algorithm uses the benchmark data set of analyzing results and the influence weight of benchmark data sets for target scale data sets to derived the association rules behind object scale data set, realize knowledge across scales derived and provide the possibility for multi-scale decision.**

Keywords-Scale Space; Big Data; Frequent Item-set; Association Rules;

I. INTRODUCTION

Multi scale is a common phenomenon in the objective world, in recent years has been widespread concern in academic circles, and gradually developed into an independent research topic - multi-scale science. Scholars of mathematics, physics, chemistry and other fields have been introduced the multi-scale theory into their own discipline and carry out a series of related research [1]. The rapid development of data fusion technology [2] and the associated model [3] application has greatly promoted the research of multi-scale domains; the collection, transmission, synthesis, filtering, correlation and synthesis of multiple information sources greatly reduce the time consumption of the scale conversion, and improve the accuracy of the data results.

Aiming at the multi-scale research on the field of data analysis, [4] using inherent multi-scale characteristic and concept hierarchy of spatial data, proposed point-object association rules analysis algorithm oriented multi-scale spatial, realized the scaling up of association rules in spatial data. [5] In conjunction with scale conversion mechanism of geography science field, proposed multi-scale clustering analysis algorithm based on enhanced weighted vector. [6] uses hierarchical multi-scale combination classification method, proposed an adaptive multi-scale segmentation combined classification algorithms, reduces the training time and get a better performance of the classifier. For the problem of the unstable of long range dependence in the existing self-similar network, [7] built network flow model and flow prediction based on discrete wavelet transform multi-scale analysis. Based on Least mean fast Wavelet Transform, considering the characteristics of the input flow and redundant wavelet transform, [8] carry out multi-scale analyzing and forecasting for self-similar network traffic. Derived from the improved Mahalanobis distance measurement to conduct fast multi-scale clustering, [9] realized data segmentation model of electronic back scatter diffraction through the presentation of data quaternion. [10] Generate an image map and track information table by scanning multi-scale data, and conduct analysis for frequent pattern and multi-scale events in the type of Depth-First-Search. In [11], an iterative and interactive hierarchical multi-scale classifier is proposed to realize the multi-scale segmentation of remote sensing images, and the classification accuracy is improved compared with the general segmentation algorithm; Based on conditional random field, the context classification of multi temporal and spatial scales of geo optical remote sensing images with different periods and resolutions was realized.

With the arrival of the era of big data, the research on the application of data analyzing is more and more in-depth, and basically committed to exploring new technologies, new methods in efficiency and accuracy to achieve across. The research of multi-scale data analysis has just started, and it is very important and urgent to explore the multi-scale data analysis method which can deal with massive data, so as to support multi-scale decision making and improve the efficiency and accuracy.

II. THEORETICAL FOUNDATION

The task of Multi-scale data analyzing is to use scaling mechanism to anti-conduct the analysis results from benchmark scale data set to other target scale data set. This section will detail describe the scaling mechanism, namely the theoretical foundation of knowledge scaling [16-19].

Scaling method most commonly used method is based on spatial statistics method, to estimate the statistical law is an optimization technique, the basic assumption is built on the spatial correlation of the prior model[19, 20]. The main idea of the relevant parameters related parameters sampling points are closer than distant sampling points are more similar, and the same degree of similarity or the size of its random spatial covariance, you can be certain lag by comparing the distance separated different values of the variables and multiple scales of measurement variability regionalized random variables to determine, and therefore become a good solution for scale conversion[21-23].

A. *The general essence of multi scale data analysis*

Kriging interpolation method is based on the structural information sampling data reflect regional variables (variogram and covariance functions provided), based on limited data sampling point to be estimated in the neighborhood of the point or block segment, consider sample points positional relationship between space, estimated to be the point of spatial relationships and treat an estimated optimal point unbiased estimate, and gives the estimation accuracy[24-26]. Because of different purposes and conditions, have produced a variety of Kriging method: to meet the steady (or intrinsic) the assumption of second order, the use of ordinary Kriging method; in non-stationary phenomenon, the available pan-g Rigby law; in the calculation of recoverable reserves, to use non-linear estimator, you can use disjunctive kriging; when regionalized variables obey a logarithmic normal distribution, the number of available kriging; when comparing data little, small distribution rules, but it does not require the estimation accuracy of available random Kriging method[27].

Kriging is the essence of valuation to be estimated by weighted summation point near point, so that the core of the weighting coefficient λ is determined. Spatial data refers to data indicating the location of spatial entities, shape, size and distribution of many aspects of information, it has a characteristic time scale, spatial scale relations. Mutual positional relationship between the data space, and have some distribution[28]. Therefore, the use of Kriging spatial scale data push, push down is very appropriate[29].

For general data set, in theory it does not have significant spatial and time scales characteristic[30]. Because most used for classification, clustering, association rules or data set size does not attribute value of spatial data, location information of other data analysis methods. In the traditional sense of space or time scale refers to units in the study of an object or phenomenon employed. However, with further research, we found that the original definition of the scale is not accurate enough or not comprehensive enough. This paper proposes the use of the concept of hierarchical distinction between the size of the scale, in fact, it extends the concept of scale, namely the concept of layering concept has partial order can be considered to be included in scale, such as: the composition of the administrative structure of the school (school, college, professional, department, class) and the like. According to the concept of hierarchical knowledge, we can assume that any data set is a multi-scale data set of mathematical partial order. One of the most exceptional cases are juxtaposed relationship between the concept of hierarchical scale that contains the relationship between the presence of any scale are not. Strictly speaking, multi-scale of some of the data does not have practical significance, that is, in practical applications, multi-scale data are generally not of practical significance and practical significance. Through the above analysis, we can draw the essence of the scale is the size of a unit of measurement range covered. Although the scale of the performance of spatial data in time and space in the areas, in general, the performance data for other category scale, but their principles and nature has not changed. That objective scale data is weighted summation of benchmarking scale data. Therefore, after making the appropriate data set general concept hierarchy to form a multi-scale data sets generated between the sample data structure information regionalized variables, according to the distribution of data, select the appropriate kriging, thus achieving the general scaling data.

B. *The typical scaling up and scaling down method*

The most typical scaling up and scaling down are respectively region ordinary Kriging method and point ordinary Kriging method, we are collectively referred to as the Kriging method. Kriging method is spatial local estimation method established on the basis of variation function theory and structural analysis, and an unbiased optimal estimation for regionalization variable aggregation in limited area. This method first defines a linear estimator:

$$P_o^*(x) = \sum_{j=1}^{m+1} \lambda_j Y(x_j) \quad (1)$$

Where, $Y(x_j)$ is the data of example point, $P_o^*(x)$ is to be estimated, λ_j is the weight for every example point, and $\sum_{j=1}^{m+1} \lambda_j = 1$. For any estimation, there exist a deviation between real value and the estimated value. $P_o^*(x)$ is a linear unbiased estimates of the optimal to the actually true value $P_o\ (x)$. Formula (1) is called as Kriging equation. Kriging coefficient λ can be expressed in the form of the following matrix multiplication:

$$\lambda = R^{-1}E \quad (2)$$

Following respectively introduce some point ordinary Kriging method and region ordinary Kriging method, and application mode in scaling up method and scaling down method. These foundations are the theoretical basis of the multi-scale data analysis.

1) Theoretical basis of scaling down analysis algorithm: point ordinary Kriging method

Using point ordinary Kriging method to implement scaling down, it is a key to determine the Kriging coefficient. E_{ij} is the covariance between the example point c and example point d of meta scale S in the K matrix. $f(x_i,x)$ is the covariance between sample point i in the meta scale S and target scale S' to be estimated. The detail form of E and F is as follows:

$$E=\begin{bmatrix} e_{11} & e_{12} & \dots & e_{1n} & 1 \\ e_{21} & e_{22} & \dots & e_{2n} & 1 \\ \dots & \dots & \dots & \dots & \dots \\ e_{n1} & e_{n2} & \dots & e_{nn} & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix} \quad (3)$$

$$F=\begin{bmatrix} e(x_1,x) \\ e(x_2,x) \\ \dots \\ e(x_n,x) \\ 1 \end{bmatrix} \quad (4)$$

Then, the specific calculation formula of weight matrix in the scaling down analysis algorithms is as follows:

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_n \\ -y \end{bmatrix} = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1n} & 1 \\ e_{21} & e_{22} & \dots & e_{2n} & 1 \\ \dots & \dots & \dots & \dots & \dots \\ e_{n1} & e_{n2} & \dots & e_{nn} & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix}^{-1} * \begin{bmatrix} e(x_1,x) \\ e(x_2,x) \\ \dots \\ e(x_n,x) \\ 1 \end{bmatrix} \quad (5)$$

2) Theoretical basis of scaling up analysis algorithm: domain ordinary Kriging method

Using domain ordinary Kriging method to realize scaling up, also need to determine Kriging coefficient. In this way, determine the matrix E is the same as the scaling up method in the point ordinary Kriging, that is to determine E by the covariance between each sample point from S' in the meta scale; F is determined in a completely different way, but by the impact factor to be estimated for target scale in each example point in the meta scale S'. We call this impact factor as the Influence Information, as shown in formula (6):

$$F=\begin{bmatrix} information_{10} \\ information_{20} \\ ... \\ information_{n0} \\ 1 \end{bmatrix} \quad (6)$$

The formula for determine Kriging coefficient by using the method of domain kriging method is shown as in formula (7):

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ ... \\ \lambda_n \\ -y \end{bmatrix} = \begin{bmatrix} e_{11} & e_{12} & ... & e_{1n} & 1 \\ e_{21} & e_{22} & ... & e_{2n} & 1 \\ ... & ... & ... & ... & ... \\ e_{n1} & e_{n2} & ... & e_{nn} & 1 \\ 1 & 1 & ... & 1 & 0 \end{bmatrix}^{-1} * = \begin{bmatrix} information_{10} \\ information_{20} \\ ... \\ information_{n0} \\ 1 \end{bmatrix} \quad (7)$$

Multi scale association rule analysis

We take domain Kriging method and point Kriging method as theoretical foundation of scaling up and scaling down, apply multi-scale data analysis algorithm framework in association rules analysis, and propose Association Rules Analysis Multi scale algorithm.

Multi scale data analysis algorithm framework can be summarized as three key steps:

1. Determine the benchmark BenchScale, process benchmark data sets;
2. Compared to the target scale TargetScale and benchmark Bench Scale, determined multi-scale analysis task of analysis direction, namely is to judge scaling up analysis or scaling down analysis;
3. Knowledge in the corresponding direction so the multi-scale transformation derived target scale data set behind the knowledge. The core of the algorithm is the knowledge of multi scale conversion, above has introduced in detail the scaling mechanism based on the theory, the key is to determine the Kriging equation in the weight coefficient matrix lambda, that is to say, whatever scale or scales push, we need to first determine the associated covariance matrix K benchmark data sets, and for D matrix to determine the calculation method in D for benchmark data set of object scale data set of influence factors matrix; calculated method D is a benchmark data set and target scale data set between the covariance matrix.

Multi scale data analysis algorithm framework is based on multi-scale association rule analysis algorithm. First clear association rules analysis in the target knowledge for association rules and analysis association rules in the process of time and computational resources consumed mainly in frequent item set analysis, so multi-scale association rules analysis algorithm to solve the core problem is the benchmark data set of frequent item sets derived object scale data set of frequent item sets and frequent item sets multi-scale conversion, and wrong target scale data set analysis.

So, the basic idea of multi-scale algorithm for analysis association rules is: first, clear benchmark scale BenchmarkScale, suitable frequent item set analysis algorithm for analysis benchmark data sets $DataSet_{BenchmarkScale}$ here can choose any efficient frequent item set analysis algorithm; then clear target scale $Scale_{target}$ and the scale of target data set $dataset_{scaletarget}$, determine the scale of association rule analysis direction; finally, the calculation of the benchmark data set analysis $dataset_{BenchmarkScale}$ results of object scale data set $dataset_{Scale}$ knowledge of weight coefficient matrix lambda, where knowledge is frequent item sets, lambda for estimating target scale data from frequent item sets of key parameters, support. Here, we calculate the weight coefficient of standard Kriging method do some improvement: K will be designated as a benchmark data set between the similarity matrix instead of the covariance matrix, can more accurately the reaction benchmark data set between similar and correlation; push process, the D matrix specified as benchmark data set of object scale data set of impact factor matrix, in pushdown processes, matrix D specified on the grounds of domain knowledge to determine the benchmark data sets with target scale data set between the similarity matrix. The problem description and basic steps of multi scale association rule analysis algorithm are as follows:

Problem Description: data set DataSet, target scale of data set $datasets_{caletarget}$, hoping to get the target scale data meet minimum support degree and minimum reliability association rules by multi dimension association rules analysis.

Basic steps of algorithms:

1. select the benchmark scale $Benchmark_{Scale}$, determine the benchmark scale data set $dataset_{BenchmarkScale}$. Benefitting analysis as the principle, according to the available computing resources, select the scale which can present the maximum utility of the existing computing resources as the scale of the benchmark.
2. Analysis all benchmark data sets by the minimum support degree, obtain frequent item set, and take the number of frequent item set union as the frequent item set candidate set of target scale data set. This candidate item set can greatest reflects the circumstances which target scale data implied frequent item sets;
3. Due to either scaling up or scaling down, it is necessary to determine the similarity matrix between benchmark data sets, so this step is to calculate the similarity between benchmark data sets; from a statistical point of view, frequent item sets is a statistical result of data set on behalf of the distribution of data set with the characteristics of itself; in statistics, similarity

coefficient is mainly used to compare the similarity and dispersion in the finite sample set. The coefficient between the finite set is equal to the intersection of two sets of elements contained in the number and the number of elements contained in the union ratio, such as shown in formula (1). The actual study researchers have applied the coefficient to data analysis. This paper uses the similarity which is between benchmark scale data sets and frequent item sets to estimate the similarity between the original data sets.

$$Similarity(C,D)=|C\cap D|/|C\cup D| \quad (1)$$

Using formula (2) computing similarity coefficient:

$$M_{mn}=Similarity(Frequentitem_m, Frequentitem_n)$$
$$=|Frequentitem_m\cap Frequentitem_n|$$
$$/|Frequentitem_m\cup Frequentitem_n| \quad (2)$$

4. Determine the analysis direction. The impact factor is benchmark dataset on the upper scale the data on the amount of data on the ratio reflects the influence degree, tendency of upper scale data sets such as division of population data of a certain area in national scale, while the Han population accounts for the vast majority of the region's population, then the "national" benchmark dataset of Chinese Population data on the overall population accounted for relatively large, equal in Han population in the region to the influence degree of the population as a whole in the analysis of population data, and Han population showed the characteristic of the data also reflects the overall population data trend. Bunches matrix elements mainly based quasi scale data set in the upper scale data quantity accounted for ratio, pushed down elements is both the similarity coefficient;
5. Determine weight matrix of the benchmark scale data sets to the target scale datasets;
6. Screening final frequent item sets of goal scale data sets and generate association rules. Final estimated frequent item sets support for all candidates in the same set of minimum support comparison, select frequent item set is not less than the set composed and produced in accordance with the minimum confidence association rules, pseudo codes for this algorithm are as follows:

III. ALGORITHM ANALYSIS

Compared with the classical data analysis algorithms, the advantage is that the proposed algorithm is only a reference scale analysis data sets. You can get a number of different levels of scale datasets behind implicit association rules. If you use the classic data analysis algorithms to achieve the same effect, you need multiple scales respectively data sets analysis, namely multiple analysis, which will cause a huge time and space overhead.

In addition, the proposed algorithm itself enforcement mechanisms to deal with the first reference scale data sets, and then the results of the multi-scale analysis are derived, if the idea of parallel computing applied this algorithm, the number of reference scale analysis parallel data sets, and then collect the results of parallel processing scale up or scale down be converted and derivation, will have a more significant efficiency has increased significantly. This large data processing is very useful. Thus, the proposed algorithm is very suitable for parallel computing, parallel multi-scale data analysis to solve program is feasible and practical significance.

The multi-scale characteristic refers to the data set itself some attribute or attribute set is related to geographical space, or time, or other relative size and size range, said clear scale meaning. In fact, the main research of multi-scale data analysis is a multi-scale and multi scale data to achieve the conversion of knowledge. Based on the theory of multi-scale data reached a preliminary multi-scale data to achieve this goal, from this perspective, scale process with multi-scale characteristic data sets has a clear basis for partitioning results in the data set also has a clear meaning of the scale; and for similar IBMT10I4D100K data set so that the multi-scale characteristic is not very obviously the data set is divided into sequential scale, process and result of meaning is not very clear. We use multi-scale data analysis algorithm framework and specific association Rule analysis and realize the conversion of knowledge at multiple scales, from the point of view of algorithm analysis, algorithm implemented with the multi-scale characteristic of the data set, no matter from the process or results, both have more practical significance, especially the need for multiple criteria decision. Therefore, whether in theory or in practical algorithms, the theory of multi-scale data and multi-scale data analysis algorithm are more suitable to have multi scale characteristics of the data set.

IV. EXPERIMENTS

In this paper, the feasibility, accuracy and efficiency of the ARAMS algorithm are verified by using the Z province data set and IBMT10I4D100K data set. Z province full population data set is a complete record of the population management and household registration and other spatial attribute information; geographical attributes can form the concept of stratification, has a good multi scale characteristics. The running environment of the experiment is Intel i7 CPU 3. 40GHz, 8G memory, Windows 8 operating system, ORACLE 11g database system, using Octave implementation algorithm. This paper uses the ARAMS algorithm to analysis the benchmark data set by adopting the classical Apriori algorithm.

Compared with the experimental results of population data set figure 1 and Figure 2, it can be found that the accuracy of the ARAMS algorithm for scaling up operation is better than the scaling down, but the accuracy of the scaling up and scaling down parts are relatively high. From Figure 1 (a), (b) shows that the upper part of the coverage and accuracy are more than 90%, and in some cases even reached 100%. From Figure 2 (a), (b) reflected in the coverage and accuracy is slightly worse than on the scaling up, but also is more than 80%. Above experiment results verify the accuracy and feasibility of the algorithm. It shows that the ARAMS algorithm is able to obtain the real frequent itemsets of target scale data set from the frequent item sets contained in the benchmark data set. There is a steep drop phenomenon in Figure 1 (a), (b) when the support is about 20%. It is due to the Y shaft size is very fine, so the subtle changes of y value in the figure will show a large change. In fact, it only

changed 5%. Figure 2 (b) shows that with the increase of the minimum support degree, the accuracy is on the upward trend, which is caused by the decrease of the proportion of false positive and false negative term sets in the overall result of frequent items. In Figure 1 (c) and Figure 2 (c) show that the average estimation error is low, especially on the experimental results of the scaling up algorithm, illustrate that the ARAMS algorithm has a good performance in the estimation of the support. In the aspect of execution efficiency, we can see from figure 1 (d), the advantage of proposed algorithms is more obvious than Apriori. From Figure 2 (d), it can be seen that the proposed algorithm has higher efficiency.

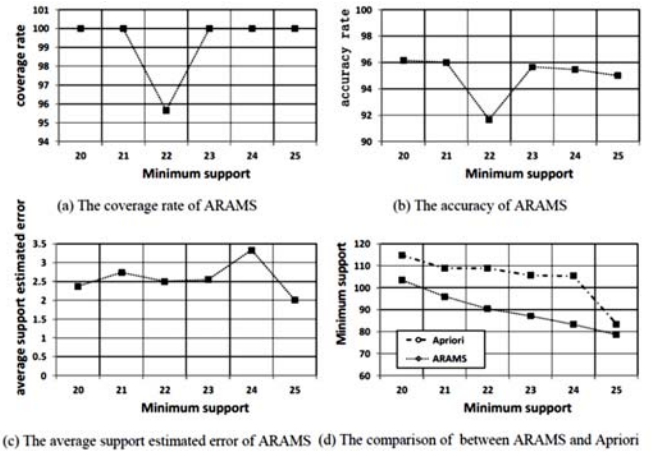

Figure 1. The scaling up experimental results of ARAMS

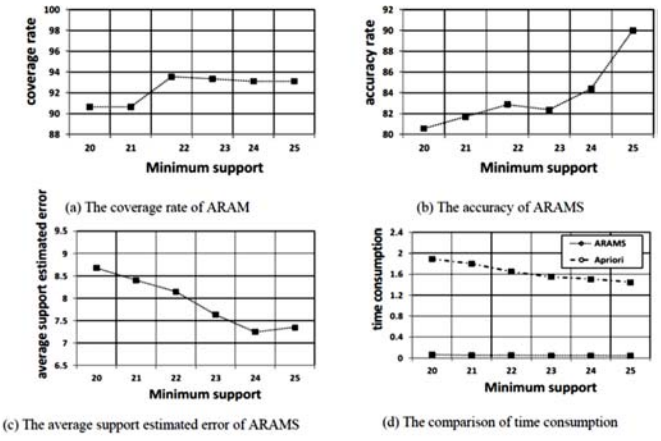

Figure 2. The scaling down experimental results of ARAMS

V. CONCLUSION

This paper introduces the basic idea of multi-scale space theory to the field of data analysis. This paper proposes data scale partitioning and data scale definition based on concept hierarchy, present the relationship between the upper and lower of scale data, and establish the theory foundation for the multi scale data space. This paper also proposes the definition and classification for multi scale data analysis, proposes algorithm framework for multi scale data analysis and present the association rule analysis algorithm for multi scale big data. The algorithm uses benchmark data sets analysis result and the influence weight of benchmark scale data set to the target scale data set to derive the association rules behind object scale data set, realize knowledge across scales derived and provides the possibility for multiple criteria decision. Finally, we conduct experiments to verify the accuracy and efficiency of the proposed algorithm. The experimental results show that the algorithm has a high coverage and accuracy, and has a low support degree estimation error, and the efficiency is higher than the traditional Apriori algorithm.

The next step, we will focus on the following aspects: research on multi-scale data analysis scaling up and scaling down, analysis the variation rule of cumulative error and accuracy in the case of cross scale; research more perfect multi-scale data theory system; multi-scale data analysis algorithm framework is applied in such as classification, clustering and other data analysis application, explore the multiscale classification and clustering analysis algorithm, and improve these experiments of the algorithms; further improve the coverage rate, accuracy degree and efficiency of the multiscale association rule analysis algorithm in the practical application, also explore better weight coefficient calculation, reduces the estimation error of support degree from theory and practice.

ACKNOWLEDGMENT

This research is supported in part by Zhejiang Provincial Natural Science Foundation of China (No. LY16F020012), Ningbo Key Laboratory of intelligent home appliances (No. 2016A22008), Key Research Center of Philosophy and Social Science of Zhejiang Province-Modern Port Service Industry and Creative Culture Research Center, Zhejiang Province Public Technology Research and Industrial Project (No. 2015C32124), Projects in Science and Technique of Ningbo Municipal (No. 2014C10047 and No. 2012B82003), National Natural Science Foundation of China (No.61300202) and Natural Science Foundation of Shanghai (No. 13ZR1452900). Zheng Xu is the corresponding author.

REFERENCES

[1] Belkin, M. and P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation, 2003. 15(6): p. 1373-1396.

[2] Hu, C., et al., Semantic link network-based model for organizing multimedia big data. Emerging Topics in Computing, IEEE Transactions on, 2014. 2(3): p. 376-387.

[3] Xu, Z., et al., Semantic based representing and organizing surveillance big data using video structural description technology. Journal of Systems and Software, 2015. 102: p. 217-225.

[4] Bell, R.M. and Y. Koren, Lessons from the Netflix prize challenge. ACM SIGKDD Explorations Newsletter, 2007. 9(2): p. 75-79.

[5] Hu, C., et al., Video structural description technology for the new generation video surveillance systems. Frontiers of Computer Science, 2015. 9(6): p. 980-989.

[6] Xu, Z., et al., Crowdsourcing based social media data analysis of urban emergency events. Multimedia Tools and Applications, 2015: p. 1-18.

[7] Bellazzi, R., et al., Temporal data mining for the quality assessment of hemodialysis services. Artificial intelligence in medicine, 2005. 34(1): p. 25-39.

[8] Luo, X., et al., Building association link network for semantic link on web resources. Automation Science and Engineering, IEEE Transactions on, 2011. 8(3): p. 482-494.

[9] Xu, Z., et al., Crowdsourcing based description of urban emergency events using social media big data. 2016.

[10] Blondel, V.D., et al., Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008. 2008(10): p. P10008.

[11] Xu, Z., et al., Generating temporal semantic context of concepts using web search engines. Journal of Network and Computer Applications, 2014. 43: p. 42-55.

[12] Xu, Z., et al., Semantic enhanced cloud environment for surveillance data management using video structural description. Computing, 2016. 98(1-2): p. 35-54.

[13] Burnett, C. and T. Blaschke, A multi-scale segmentation/object relationship modelling methodology for landscape analysis. Ecological modelling, 2003. 168(3): p. 233-249.

[14] Ding, H., et al., Querying and mining of time series data: experimental comparison of representations and distance measures. Proceedings of the VLDB Endowment, 2008. 1(2): p. 1542-1552.

[15] Džeroski, S., Multi-relational data mining: an introduction. ACM SIGKDD Explorations Newsletter, 2003. 5(1): p. 1-16.

[16] Huck, K.A. and A.D. Malony. Perfexplorer: A performance data mining framework for large-scale parallel computing. in Proceedings of the 2005 ACM/IEEE conference on Supercomputing. 2005. IEEE Computer Society.

[17] Jenatton, R., et al. Multi-scale mining of fMRI data with hierarchical structured sparsity. in Pattern Recognition in NeuroImaging (PRNI), 2011 International Workshop on. 2011. IEEE.

[18] Khan, S.S. and A. Ahmad, Cluster center initialization algorithm for K-means clustering. Pattern recognition letters, 2004. 25(11): p. 1293-1302.

[19] Knobbe, A., et al., Multi-relational data mining. 1999.

[20] Xu, Z., et al., Knowle: a semantic link network based system for organizing large scale online news events. Future Generation Computer Systems, 2015. 43: p. 40-50.

[21] Kopanas, I., N.M. Avouris, and S. Daskalaki, The role of domain knowledge in a large scale data mining project, in Methods and Applications of Artificial Intelligence. 2002, Springer. p. 288-299.

[22] Lancaster, A., et al. PyPop: a software framework for population genomics: analyzing large-scale multi-locus genotype data. in Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. 2003. NIH Public Access.

[23] Li, S.-T., S.-W. Chou, and J.-J. Pan. Multi-resolution spatio-temporal data mining for the study of air pollutant regionalization. in System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on. 2000. IEEE.

[24] Li, S.-T. and L.-Y. Shue, Data mining to aid policy making in air pollution management. Expert Systems with Applications, 2004. 27(3): p. 331-340.

[25] Low, Y., et al., Distributed GraphLab: a framework for machine learning and data mining in the cloud. Proceedings of the VLDB Endowment, 2012. 5(8): p. 716-727.

[26] Machiraju, R., et al., EVITA—Efficient Visualization and Interrogation of Tera-Scale Data, in Data mining for scientific and engineering applications. 2001, Springer. p. 257-279.

[27] Mennis, J. and D. Guo, Spatial data mining and geographic knowledge discovery—An introduction. Computers, Environment and Urban Systems, 2009. 33(6): p. 403-408.

[28] Streilein, W., et al. Fused multi-sensor image mining for feature foundation data. in Information Fusion, 2000. FUSION 2000. Proceedings of the Third International Conference on. 2000. IEEE.

[29] Tsytsarau, M. and T. Palpanas, Survey on mining subjective data on the web. Data Mining and Knowledge Discovery, 2012. 24(3): p. 478-514.

[30] Vespignani, A., Predicting the behavior of techno-social systems. Science, 2009. 325(5939): p. 425.