International Journal of Database Theory and Application
Vol.9, No.12 (2016), pp.79-88
http://dx.doi.org/10.14257/ijdta.2016.9.12.08

Green Mining Algorithm for Big Data Based on Random Matrix

Wang Canwei

Department of Information and Engineering ,
Shandong Management University, Jinan 250357,China
wangcanwei@sina.com

Abstract

Due to big data with related multi-dimensional characteristics, the effective means how to build processing mechanisms and algorithms are still problems; so that the algorithms on big data processing huge resources and time cost of computing, resulting in wasting of energy; for this problem the present study proposes a large data processing algorithm of random matrix theory application, can effectively improve the processing efficiency, thereby increasing the utilization of energy. Results show that the proposed algorithm can effectively reduce the amount of calculation, thus saving and calculating the required energy.

Keywords*: energy efficiency; random matrix; large data; convergence*

1. Introduction

With the rapid development of the Internet, mobile Internet, Internet of Things, cloud computing and other information technology, information technology and human aspects of social integration of the cross, gave birth to transcend any era of massive data, big data era has arrived. Big Data is considered to be the new "oil" of the information age, the large data tremendous value if good use, will bring revolutionary development in many areas. Thus, large data field of research has attracted wide attention of industry, academia and government. Nature magazine in 2008 launched the first large data special issue devoted to the massive amounts of data for all aspects of the Internet, economic, environmental and biological effects of influences and challenges. Science magazine launched in 2011 monograph major issues surrounding the scientific research big data discussion illustrates the importance of big data for scientific research. In 2012, the Obama administration released data research and development programs that "By collecting, handling large and complex data and gain knowledge and insights to enhance the ability to accelerate the pace of innovation in science, engineering, and strengthen US national security stop , transform education and learning." For six federal agencies on the mouth and announced plans to invest $ 200 million to improve the collection, storage, retention, manage, analyze and share massive amounts of data and advanced core technology. Big data with four basic characteristics, namely a huge body mass, the variety and timeliness of high and low density values. These data determine the characteristics of large research fast and effective analysis of large data processing technology urgency and necessity; therefore, large data analysis has been the core of research in the field of big data. Big data analysis is the concept of big data and methods, that massive, diverse, fast-growing, low-value density data analysis, and is found to assist in decision-making hidden mode, unknown related process relations and other valuable information. Large-scale data [1-3], the existing data mining or statistical analysis of the data is often simplified and abstract, which will hide the real structure of the data set, and the data visualization. You can restore and even enhance the data in the global structure and details. In the big data environment, a large data itself features for data visualization made more urgent needs and tougher challenges. In this study, temporal and spatial big data

ISSN: 2005-4270 IJDTA

environments multidimensional data visualization problem is studied, comprehensive utilization of large data processing and data visualization technology to improve data visualization larger data size when the big data is a data volume surge (from the beginning of the ERP/CRM data amount of data, and gradually expand to increased Internet data, to the things of sensor data and other relevant information), and the complexity of the data improved. Large-scale data can be said that a qualitative change is the amount accumulated to a certain extent after the formation of the data type of a large variety of data, both there are structured like the original database data and other information, but also text, video and other unstructured information, and data acquisition and processing speed requirements are also increasing faster.

With the advent of the era of big data, big data analysis and processing technology becomes increasingly important. Big Data technology development was mainly due to the development of parallel computing hardware and software technology, the industry also benefited from the rapid growth in recent years of large data processing needs. Big data processing technologies including cloud computing, parallel computing technology, distributed file systems, distributed databases. Cloud computing technology is the basis of large data processing, and it is big data analysis techniques support. Cloud computing services can generally be divided into three levels: infrastructure as a service layer, platform as a service layer and software as a service layer. Cloud computing technology is the core principle of big data analysis and processing technology, which is also large data base platform analytical applications. Parallel computing technology is the core technology of large data processing. With the expansion of the scale of the problem of data volume growth and calculation, traditional computing methods cannot meet the actual needs of the application of computing power and computing speed, so large data processing take way of parallel computing. Currently has developed a variety of computing framework having parallel processing automation can provide, such as Google MapReduce and Hadoop MapReduce parallel computing framework, as well as memory-based computing, big data can provide a variety of computing models Spark systems. In the large-scale cluster environment, how to solve large-scale data storage, access management and other issues, is a major problem in large data processing need to be resolved. Distributed file systems aims to provide a good solution. Distributed database is the use of high-speed computer network to a data storage unit connected to physically discrete uniform composition of a logical database. With the advent of the era of big data, distributed database technology has been rapid development, based on a distributed relational database data model to retain under the traditional database and basic features, from a centralized to a distributed memory storage. Some technologies may harm the environment and cause a lot of negative issues, such as pollution, waste, resource depletion and ecological disruption. Therefore, by effectively utilizing big data can address issues related to the environment. Renewable resources such as water, due to limitations of the prior art, is the consumption or depletion rate faster than they can be self-renewal rate. Big data analysis support to reduce resource consumption. In recent years, the European Union launched a research project to improve water management. Through the analysis of large data, so that the project emphasizes energy efficiency is the pumping station as possible and reduce the number of leaks in the water supply network[4-6].

2. Related Works

2.1. Big Data Research Situation

Multi-dimensional data refers to the data having a plurality of variable dimension attributes, multi-dimensional data in real life can be seen everywhere and is important, people often make decisions based on multi-dimensional data analysis. When the amount

of data is not large and the dimensions of the data are not high, individuals can more easily make decisions based on data analysis. While when the data dimension increases, large amount of data, we need to rely on help. Thus, the multi-dimensional visualization of data in data analysis has broader applications. The purpose of multi-dimensional data visualization is to explore the distribution and pattern of multi-dimensional data items, and uncover the relationship between the properties of different dimensions. At present, foreign scholars have proposed a variety of multi-dimensional data visualization methods. These methods depending on their visual principle can be divided into geometry-based technology, pixel-oriented technology, icon-based technology, based on level technology, graphics-based technologies and based on dimensionality reduction mapping technology. Among them, the geometry-based technology is the most commonly used method. The basic idea of the technology that based on the geometry of multi-dimensional visualization is the way on which geometric painting or geometric projection of high-dimensional data is mapped to the low-dimensional space, such as a point or a line to represent multi-dimensional information object, but the data do not apply to more dimensions, so the data set is relatively easy to observe the distribution of multi-dimensional data and found the different points. Based on the geometry of multi-dimensional visualization techniques include parallel coordinates, coordinate radiation, scatter and scatter matrix and the like. The research and application of parallel coordinate is most widely used in multi-dimensional visualization techniques. In traditional data visualization methods[7-8], the axes perpendicular to each other, each data object corresponding to a coordinate of a point. The parallel coordinates visualization methods used a plurality of mutually parallel axes, each coordinate axis represents a dimension attribute of data, and each data object corresponds to a polyline through all axes. Parallel coordinate dimensions can display more data in two-dimensional space, it can not only reveal the distribution of data on each property, but also the relationship between the two properties can be described adjacent. However, parallel coordinate data expression dimension depending on the screen width, when the number of dimensions too is large, it will make the horizontal distance between the axis becomes small, resulting in difficulty to identify the data structures and relationships. And in big data environments, Parallel Coordinates will appear dense lines with overlapping coverage problems. Related scholars according to the line gathered characteristics proposed the clustered parallel coordinates visualization method.

Scatter plot is another commonly used multi-dimensional visualization method. Its essence is the abstract data objects mapped to a Cartesian coordinate system in two-dimensional space. Data object position coordinate system reflects its distribution characteristics, can be intuitive, and effectively reveal the relationship between the two properties. Dimensional scatter plot can show limited multi-dimensional data, the method can be used scatter plot matrix. Scatter plot matrix dimensions variable combinations of two multi-dimensional data matrix as a panel, each panel corresponding to the variable plotted in two dimensions, thereby completing the multi-dimensional mapping of the two-dimensional space. Meanwhile, the scatter plot intuitive and widely used, scatter plot matrix can be more easily accepted by the user, and can be very effective in revealing the associated attributes. However, when data dimensions increase the number of scatter plot matrix panel will rapidly increase, showing entirely in a limited screen space.

Pixel-oriented multi-dimensional visualization technology, the basic idea in accordance with the dimension of the data with high-dimensional space is divided into a plurality of sub-windows. Each sub-window corresponding to one-dimensional data respectively, with the color of the pixel to represent corresponding dimension values. Some scholars have proposed a method that use intensive display pixels of different colors to cube expression of the data stored in large databases. Multi-dimensional data points for each series is represented by a rectangle of pixels, each pixel represents an attribute dimension, color-coded data value and rectangles arranged according to certain layout strategy in

two-dimensional space, generating an entire block of pixels. Pixel-oriented visualization technology using recursive model, spiral model, the circumferential segmentation models and other methods distributes data. Its purpose is to show as much data as possible on the screen window for visualization of large data sets. Based on the multi-dimensional visualization technology icon, the basic idea is to use an icon with a plurality of visual features to express multi-dimensional information, a visual icon of each feature can be used to represent one-dimensional among multi-dimensional information. It is applicable to small dimension that the data sets are special meaning sets for some dimensions. Users can more accurately understand the meaning of these dimensions based on icons.

Based on hierarchical multi-dimensional visualization technology, multi-dimensional space is divided into several sub-spaces, and each sub-space hierarchically are organized and graphically represented, most of which use the tree structure. Based axis mapping multi-dimensional visualization techniques, linear or non-linear transformation multi-dimensional data projector are projected or embedded in low-dimensional space, and try to keep the data features in a multi-dimensional space unchanged for visualizing high-dimensional data set and present data overall structure and distribution.

Two important features of the multi-dimensional time-series data are timing and multi-dimensional. Some data will be showing the sequential changes of performance data over time regularly or periodically, while others have no law at all. Refers to multi-dimensional data having a plurality of attribute dimensions, with the number of the dimensions increases, the data will also become more complex. The purposes of multi-dimensional time series data visualization, is to help the experts find that the relationship among the dimensions of data for multiple properties, and explore data attributes change over time against the law or the law of a singular point.

2.2. Random Matrix Theory

Random matrix theory was proposed in the 1850s by Wigner in the study of heavy nuclei levels, the basic idea is to use a random matrix ensemble simulation of physical systems. Random matrix theory is an important mathematical tool for statistical analysis of complex systems. Through the spectrum and eigenstates complex system of statistical analysis, random matrix theory obtained the random degree of actual data, which reveals the behavior characteristics of the actual data associated with the overall.

Because the random matrix theory evolved from physics after Wigner first proposed using random matrix ensemble to establish realistic physics model, random matrix theory [9-11] (RMT) in other areas have also been rapidly developed. For example, the financial markets, radio communications and so on. Thus, the successful application of the random matrix theory in statistical physics for the development of random matrix theory provides a good opportunity. In physics, the problem that can be solved exactly is few in general. And with the increase of the complexity of the object in theoretical physics and the development of research, the problem can be solved exactly the less the less. For example, in Newton theory of gravity in the two-body problem can be solved exactly, but generally the three-body problem is not; in general relativity and even general two-body problem is difficult to solve, the problem can also be solved exactly monomer; and to the quantum field. Theory monomer is even difficult to solve the problem. In reality, many subjects are having a very complex structure, and the object of study is often a large number of multi-body structures. Traditional analytical methods "exact calculation" formula in the face of these complex systems is no longer applicable. Even to get some results, the cost is often huge.

So, people are trying to find various ways to get approximate results that we want. The statistical method is just one of the very good approximations that people like to find. So, when faced with the development of physics as the plight of using statistical methods in physics will be developed, it is called "statistical physics." Statistical physics basic concern is no longer a microscopic state, but the macroscopic state of the system. So that

it can determine the characteristics of the macro-up system. We could say that the macro is a statistical average amount corresponding microscopic amounts. Various systems of macroscopic behaviors and states use the "ensemble" to classify.

"Ensemble", refers to certain macroeconomic conditions, a large number of identical nature and structure, in various states, a collection of separate systems. In classical statistical physics, although the microscopic states of different systems are different, but any Hamiltonian system of the same ensemble are the same. However, with increasing complexity of the object of study, it was discovered that some Hamiltonian system are even unobtainable. So based on this research, Wigner studied randomized Hamiltonian system and expanded the statistical physics in the form of a matrix to represent. This opened up a new theory that was random matrix theory. Thereafter, the independence of the Dyson and others from the structure, symmetry, matrix element distribution, based transformation invariance; symmetry angles RMT conducted more in-depth study. Dyson found in nature Gauss Distribution introduce random matrix theory. And giving three types of ensembles, namely: the Gaussian unitary ensemble, Gaussian orthogonal ensemble and Gauss sympletic ensemble.

In the era of big data background, what is the new features of big data? Data acquired in the era of big data is not difficult; the difficulty lies in data processing. Big data collected to have the following three characteristics 1. Between social survey data entry uncertain relationship, which is caused by many reasons. From a sample of the object alone, the data obtained from it is reflected in the characteristics of the sample itself, there is correlation among these data. As for multiple sample objects, there are similarities, interacting with the divergence and other relationships among them. If we consider follow-up survey of the same target group for many years, then each has certain correlation between data updates. All these make the relationship between social surveys data has a fairly complex correlation. And these questions could not be found in traditional statistical survey. 2 The data is discrete. The data types include discrete variables, binary and categorical variables, which are determined by the data source. Either through self-administered questionnaire or network survey, there is no strict logical relationship between the data, what can be obtained merely play the role of labeling. 3. Ambiguity. Vagueness is a characteristic of lots of things and phenomena in the objective reality. In particular, in the verbal communication, a wide range of data sources will make one object correspond to a variety of expressions, concepts. The data distribution described by statistics can be divided into two categories: one is the number of the center position, and the other represents the number of the degree of variation (or degree of dispersion). Both complement each other, together reflect the whole picture of the data. Frequency analysis is commonly used in analysis of frequency, analysis of variance and so on. Correlation analysis is an indicator to show how closely interdependent relationship between different phenomena. The method used to measure simple linear correlation coefficient is Pearson simple correlation coefficient. The main task of regression analysis is to estimate the parameters based on sample data. In regression model, the parameters of a model were tested and judged, and then to predict. The basic idea is, on the basis of correlation analysis, the general change in the relationship between two or more variables correlate with the measured value to establish an appropriate data model to infer from a known quantity of another unknown the quantity.

In the energy sector, due to the huge amount of data that is required for effective data analysis tools to analyze the problems in the energy sector, this can make an effective energy analysis with the help of Internet decisions. Big data is an emerging pattern applied to the data set size beyond the ability of commonly used software tools to capture, manage, process data can be tolerated in the elapsed time. Various techniques are being discussed to support large data processing, such as massively parallel processing database, scalable storage system, cloud computing platforms and MapReduce. Massively parallel processing database is relevant to address tolerance through real-time operating time.

Appling of random matrix theory, massively parallel processing database mining algorithm is optimized, thereby it increases the convergence speed parallel data processing.

3. The Proposed Scheme

At each sampling time, the sampling N of the data, the data is represented as a vector, this vector can be expressed $\mathbf{x}_i$, and $\mathbf{x}_i \in R^N$, for the number of data samples referred to as T, and in fact, a huge number of numbers. And these huge sample data collection form a function that is $f(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)$. Having the characteristics of large data $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T$ is heterogeneous and diverse characteristics of the variables or parameters described as large, complex systems or subsystems. As for the dimension data representing the samples collected, the data matrix large orders $\mathbf{X} \in \square^{N \times T}$, when one of the elements to $x_{i,j}$ meet their zero mean and variance σ^2; if the variance is a finite number, and where the elements meet the characteristics of independent and identically distributed, according to the random matrix theory , as $N, T \to \infty$ is its covariance matrix $\mathbf{S} = (1/N)\mathbf{X}\mathbf{X}^H$ converge to Marchenko-Pastur law, the probability density distribution function

$$f_{\mathrm{ESD}}(\lambda_s) = \begin{cases} \dfrac{1}{2\pi\lambda c\sigma^2}\sqrt{(\mathrm{b}-\lambda)(\lambda-a)} & a \le \lambda \le b \\ 0 & other \end{cases} \tag{1}$$

Wherein $a = \sigma^2\left(1-\sqrt{c}\right)^2$, $b = \sigma^2\left(1+\sqrt{c}\right)^2$; the analysis of big data correlation time, can use this parameter to set the threshold, you can find relevant data, according to this theorem, you can easily find the relevant data to achieve retrieval compression of the data, which is the primary means of data preprocessing, convergence and efficiency in subsequent experiments can be verified.

In the calculation of cloud computing, the data related to the content of the operation, according to the status of data distribution, and now each node obtained the relevant values, partial results and then calculated transmission cost to the cloud, the cloud judged the correlation of the data, according to relevance, as determined by the cloud terminal whether your data sent to the cloud or not. Thereby reducing the transmission cost of data and reducing the amount of calculation and ultimately reduce the energy consumption to achieve a green development of the data.

Kernel methods made an important development and widely application of these methods include kernel principal component analysis, the use of nuclear techniques included support vector machine, nonlinear dimensionality reduction methods, *etc.* in machine learning and statistical analysis. Kernel method is essentially people multivariate analysis in infinite-dimensional space, through the introduction of the kernel, to embed data into an infinite-dimensional space it is a superior value characteristics: Embedding no specific data, through all the calculations, and the use of finite-dimensional kernel matrix. Nuclear Matrix closely related to the matrix plays an important role in manifold learning. In classical statistics, which has become the mainstream in spatial statistics and geo-statistics, the most obvious is the dimension data, and research applications as well as in the nuclear manifold learning methods. It is often assumed that the data have the low-dimensional manifold or low-dimensional structure. Therefore, the nuclear matrix based on the theoretical analysis and most manifold learning methods are in low-dimensional environment. Due to the importance of high-dimensional data analysis, and there are already some random kernel methods which make nuclear matrix in high-dimensional environmental could work too, and the research about it is necessary based on some data.

Suppose $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T$ subject to the following two distributions:

Norm $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T$ on four keys from the uniform distribution manifolds.

From $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T$ having a concave isotropic and logarithmic distribution.

Conditions that high-dimensional data is large:

1. So that when T and $N \to \infty$ there is a constant y $T/N \to y \in (0, \infty)$

2. When T and $N \to \infty$, $T/N \to 0$.

Set $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T$ are independent and identically distributed random variables and have the same distribution; hypothesize $\mathbf{K}_n = \left(f\left(a_{N,\mathrm{p}}{}^{-1} \mathbf{x}_i^T \mathbf{x}_j \right) \right)_{T \times T}$, and both the second-order derivative and the first-order derivative of it at 0 exists. Provided, when T and $N \to \infty$, $T/N \to y \in (0, \infty)$; the F^n weak convergence with probability 1 to distributed random variables $\left(f(1) - f(0) \right) + f'(0) S$.

As for the similarity determination of data, we need to find similarity metrics of two vectors. In the present study, the distance between the spectral measure, referred to as

$$d\left(F^{\mathbf{X}_1}, F^{\mathbf{X}_2} \right) \tag{2}$$

According to the theory of random matrices, the upper bound of the spectrum of this measure is

$$d\left(F^{\mathbf{X}_1}, F^{\mathbf{X}_2} \right) \le \sqrt{\frac{1}{T} tr\left(\mathbf{X}_1 - \mathbf{X}_2 \right)} \wedge \frac{rank\left(\mathbf{X}_1 - \mathbf{X}_2 \right)}{T} \tag{3}$$

This can be obtained if more than the upper bound of the two data were judged to be irrelevant data. Thereby this mathematical formula could be used to filter the data in general applicability big data, and different data structures will not change. So the integration of heterogeneous data has a very wide range of applications.

4. Scheduling Method Based on Cloud Computing Random Matrices

Cloud data processing platform is the core of the system of data storage and data processing. The platform structure is a distributed file system and parallel computing framework, integration interfaces to external systems, multi-dimensional analysis of massive data knowledge mining, data analysis and other results show service, as shown below. The system can support large-scale distributed data acquisition, multi-dimensional analysis of parallel and parallel data mining.

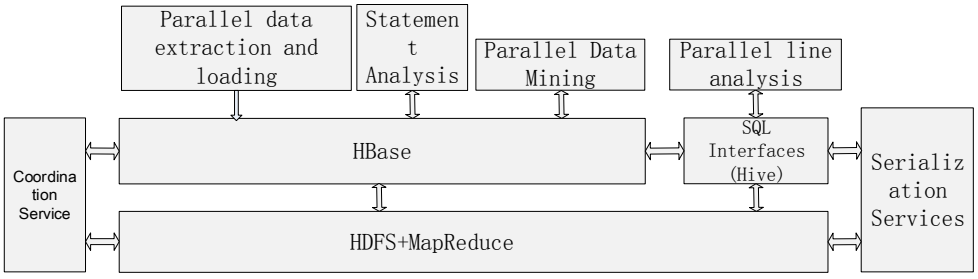

Figure 1. Data Relation

As can be seen from Figure 1, the main provider of cloud data processing platform features four categories, which are distributed data storage, parallel data extraction, transformation and loading (*i.e.* ETL), parallel data mining and multi-dimensional analysis of parallel. Wherein the distributed data storage uses HBase to store data, which

can greatly improve the efficiency of data access. Since the underlying storage HBase uses HDFS, so that the distributed data storage is safe and reliable. ETL data is used to complete the data source through the extraction, transformation, loading process into a series of HBase. Parallel line analysis makes multi-dimensional analysis using Hive, class Hive SQL interface, and provides easy interfaces to use, multi-dimensional analysis greatly simplifies the difficulty of implementation. MapReduce parallel data mining based on parallel computing framework to achieve clustering, classification, prediction, outlier analysis and other data mining algorithms making it possible to discover valuable knowledge from massive data.

Cloud data processing platform in parallel multi-dimensional analysis technology uses Hive to achieve the goal of analysis. And the specific implementation steps as follow: the first is to map data stored in HBase or HDFS to the Hive, and then type SQL interfaces(Hive), data multi-dimensional analysis and statements analysis into MapReduce job executed in parallel, enabling the entire multi-dimensional analysis of parallelization.

Suppose the collection is represented by a directed acyclic graph $W=(T,E)$, in which $T=\{t_1,t_2,\cdots,t_n\}$ is addressed on behalf of the computing tasks. And there will be a link (marked $\{t_i,t_j\}$) to any pair of tasks in the collection. If there is data exchange between node t_i ($t_i \in T$) and node t_j ($t_j \in T$) corresponding to the link $\{t_i,t_j\}$, the edge e_{ij} ($e_{ij} \in E$) will be established. This can be carried out between each other and makes resource scheduling, and task can be loaded into each other. Based on this definition, any subtask could not be calculated until the superior subtasks are completed. Additionally, each task of the workflow W have a final time, *i.e.* δ_W, in order to ensure Quality of Service and the execution time limit.

In the mobile Internet cloud computing provides a multi-service platform. Therefore, the time limit for each business processes is different. And the impact of resource scheduling on their time needs in the entire cloud computing is very large, so it is need to constantly change the topology according to the conditions, each node needs to schedule computing resources to improve business service quality. For example, if a business needs to be done within 60 seconds, we need to compare computational speed and transmission energy of each node.

Let mission time, calculate the amount needed for this computation with floating point calculations as the unit, so you can get the link between the two variables according to formula (4)

$$T_{t_i}=C_{t_i}/(P_s*\rho_s) \tag{4}$$

Automatic scheduling and allocation of computing resources has different optimization objectives, for example, some objective require some optimization goals in the final period to meet the conditions to minimize the costs of implementation. This article defines a scheduling object, which is a series of resource allocation on behalf of relief, for a computing task, the expenditure of mapping resources can be divided into two parts that are the total cost of implementation and the total execution time. Each resource represents the first task of computing resources occupied blocks.

Each particle is represented by a defined speed (or step length) and direction of the search, and the algorithm search each particle according to the local optimal solutions. In each step of the search algorithm, it can be solved in accordance with its search speed and direction. In each iteration, the position and speed of iteration can be represented by the formula (5) and (6)

$$\mathbf{x}_i(n+1)=\mathbf{x}_i(n)+\mathbf{v}_i(n) \tag{5}$$

$$\mathbf{v}_i(n+1) = \lambda \mathbf{v}_i(n) + \beta_1 r_1 \left(\mathbf{x}_i^*(n) - \mathbf{x}_i(n) \right) + \beta_2 r_2 \left(\mathbf{x}^*(n) - \mathbf{x}_i(n) \right) \quad (6)$$

5. Experiment and Analysis

In order to verify the validity and rationality of the proposed algorithm above, several experiments have been done. In the experiments, making average power consumption as the measure standard, four state-of-the-art algorithms, namely proposed scheme, SALSA, random method and minimum delay method respectively, are compared. And the polyline chart of the experiments is shown as Figure 2:

As shown in Figure 2, the average power consumption of the proposed scheme in this paper is the lowest among the four algorithms. With the arriving rate of data which is measured by Mbps unit increases, the disparity between the proposed scheme in this paper and the rest methods become greater and greater. At 0.6 point in abscissa which is the most obvious difference, the average power consumption of the minimum delay method is two times of the proposed scheme in this paper. All of the experiments fully show that the proposed scheme in this paper is better than the rest methods.

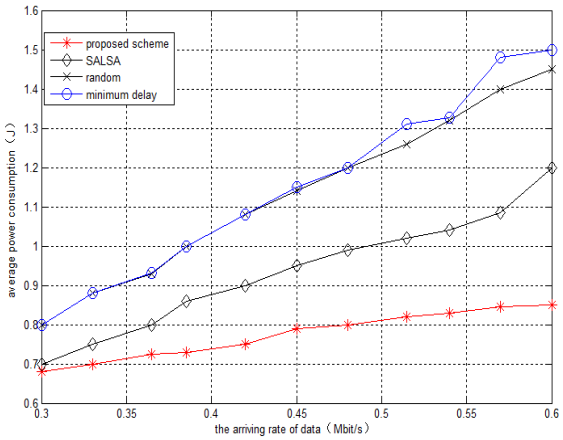

Figure 2. Power Consumption Compare

6. Conclusion

Based on the characteristics of random matrix, data screening methods and data screening threshold value were presented. And according to the characteristics of large data processing, this paper proposed a resource scheduling method with random matrix theory for large data based on cloud computing. Experiment results show that the proposed method can effectively reduce the cost of computing and communication scheduling. Due to achieve energy saving, this method is a green algorithm for large data processing.

Acknowledgements

This research was supported by Shandong province education science planning special topics of computer teaching " the application and research of cloud computing technology in computer practice teaching ", the project number: YBJ15005.

References

[1] J. Fan, F. Han and H. Liu, "Challenges of big data analysis", National science review, vol. 1, no. 2, **(2014)**, pp. 293-314.
[2] C. Zhang and R. C. Qiu, "Data modeling with large random matrices in a cognitive radio network testbed: Initial experimental demonstrations with 70 nodes", arXiv preprint arXiv: 1404.3788, **(2014)**.
[3] Y. Zhang, M. Chen and S. Mao, "Cap: Community activity prediction based on big data analysis", Network, IEEE, vol. 28, no. 4, **(2014)**, pp. 52-57.
[4] X. He, Q. Ai and J. Ni, "3D Power-map for Smart Grids---An Integration of High-dimensional Analysis and Visualization", arXiv preprint arXiv: 1503.00463, **(2015)**.
[5] G. Green and T. Richards, "Interpreting Results of Demand Estimation from Machine Learning Models", Agricultural and Applied Economics Association, **(2016)**.
[6] P. Cerchiello, P. Giudici and G. Nicola, "Big data models of bank risk contagion", University of Pavia, Department of Economics and Management, **(2016)**.
[7] A. Cichocki, "Tensor networks for big data analytics and large-scale optimization problems", arXiv preprint arXiv:1407.3124, **(2014)**.
[8] K. H. Low, J. Yu and J. Chen, "Parallel Gaussian process regression for big data: Low-rank representation meets Markov approximation", arXiv preprint arXiv:1411.4510, **(2014)**.
[9] R. Dubey, A. Gunasekaran and S. J. Childe, "The impact of big data on world-class sustainable manufacturing", The International Journal of Advanced Manufacturing Technology, **(2015)**, pp. 1-15.
[10] J. Paisley, D. Blei and M. I. Jordan, "Bayesian nonnegative matrix factorization with stochastic variational inference", Handbook of Mixed Membership Models and Their Applications. Chapman and Hall/CRC, **(2014)**.
[11] E. D. Schifano, J. Wu and C. Wang, "Online updating of statistical inference in the big data setting", arXiv preprint arXiv: 1505.06354, **(2015)**.

Author

Wang Canwei, received the bachelor's degree in science from Liaocheng University and the master's degree in engineering from Shandong Normal University in 2004 and 2012 respectively. He now is a visit scholar at department of computer science and technology of Nanjing University. He is currently research interests on Machine Learning and Big Data Analysis.

