# Research of Graph Mining Base on Cloud

Cen Li[1]

[1] Zhangjiajie College, Jishou University, Zhangjiajie, Hunan, 427000

hunter2011@foxmail.com

**Keywords:** Big Data; Data Mining; Cloud Computing; Energy Optimization; Dynamic Graph

**Abstract.** With respect to the problems of idle consumption and luxury consumption existing in the present cloud computing platform, and the current situation that the traditional graph mining method could not satisfy the massive data mining, A dynamic graph mining method for minimum energy consumption optimization cloud was proposed, resources integrated utilizing and massive data mining has been solved. First, the cloud computing energy measurement formula was proposed, and analyzes the rationality of two types of task scheduling policy theoretically. Second, one time the problems of system energy optimization and system operation efficiency was considered. Under the condition that the system operates well, it converts the problem of system energy optimization into system cost control, and the total cost of the objective function was proposed, a model for computing adaptive allocation algorithm and the minimum energy consumption optimization cloud was designed.

## Introduction

In the big data [1-2] under the background data with unprecedented speed rapid growth, which has become a mass data available, valuable, basic resources, is one of the most important government assets . Specific representation of a large variety of data, the most important and most widely used form of a diagram showing the structure, such as social networks, RFID (Radio Frequency Identification, RFID), biological genes, e-commerce, the Internet and so can be used data graph. However, with the passage of time and changes in the external environment, internal structure may also be changed, such a diagram called a dynamic graph or uncertain graph [3]. In addition, e-commerce transactions in the user data will also change with circumstances, occur when consumers and businesses to return, or for some reason when consumers replace frequented merchandising business, user behavior data structure on trading It will change occur. In these cases, how to find a common characteristic diagram data or hidden information, and access to structural variation graph with important research significance, and has become a hot research question.

## The Minimum Energy Optimization Cloud Model

**Cloud Platform Consumption Metrics.** In the cloud computing platform, the system is mainly reflected in the hardware energy consumption and energy consumption in terms of hardware energy consumption on the basis of electricity, network, air conditioning, etc. This article will not consider necessary energy, paper studies energy consumption problem, that is, by adjusting the internal operation mechanism cloud computing platform to reduce overall energy consumption. This article will use cloud computing platform to optimize task scheduling strategy to reduce energy consumption, before talking about the specific method described first task scheduling mechanism and energy metrics traditional cloud computing platform.

Task random scheduling mechanism, although simple, but does not take into account the overall system efficiency and energy optimization and other issues. Herein by reference [15] M/M/1 (M/M/1 model) queuing model to study computing platform to measure the energy consumption of the cloud, the specific implementation method based on the model of energy optimization. Firstly, measure the energy consumption of the parameters described in the following table:

Table 1 Energy consumption metrics table

| Parameters | Explanation |
|---|---|
| $i \in [1, I]$ | Calculate the number of nodes |
| $j \in [1, J]$ | The number of tasks |
| $\mu_{ij}$ | Tasks $j$ compute node $i$ service rate |
| $\rho_{ij}$ | Tasks $j$ compute node $i$ servicer intensity |
| $P_{ij}$ | Tasks $j$ compute node $i$ service rate |
| $E(\rho_i)$ | Compute nodes $i$ desired service intensity |
| $P_d$ | Idle probability of individual compute nodes |
| $E(C)_D$ | Cloud platform idle power consumption |
| $E(C)_B$ | Perform energy cloud platform |

**Minimum Energy Optimization.** Total consumption cost optimization objective function to solve the problem is a NP (non-deterministic polynomial) difficult problem, there is based on graph theory, the precise method for solving integer programming and heuristic solution method based on approximate solutions. Want NP-hard problem in a short time is difficult to solve accurately, often requires powerful hardware resources and efficient processing technology. Algorithm uses random rotation roulette with a combination of methods to select the optimal allocation scheme. Specific steps are as follows:

Algorithm 1: MECOTAA

Input: computing tasks lists, parameter settings in Table 1, the minimum cost threshold

Output: optimal task allocation scheme

1) The system initialization function $InitialOp()$ will be calculated into the task list $\boldsymbol{TL}$ task scheduling queue, according to the parameters set in Table 1 set parameters, which $J = Length(\boldsymbol{TL})$ is given by the user estimated time costs, $r = 1$ represents only one input file type.

2) According to equation (5) Design task allocation $AS = (A_{rij}, F_{rij}, X_{rij})$ , and assuming that all resource scheduling time cost of 1, which set: $f_{ri} = 1$ , $\delta^t_{rij} = 1$ , $\beta^t_{rij} = 1$ .

3) Randomly generated $(A_{rij}, F_{rij}, X_{rij})$ in the initial allocation scheme, then according to the formula (9-12) to determine the resulting program is valid, if it is valid according to the formula (5-8) $X(P)$ $Y(P)$ $Y(P)$ and $Cost(P)$ were calculated, as well, otherwise rebuild the distribution scheme.

After obtaining the best task allocation, the system allocates tasks to perform in accordance with the program will enter in order to achieve optimal energy consumption and performance. Model as shown below:

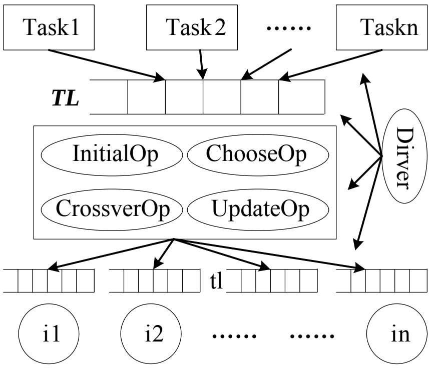

Fig. 1 Minimum energy optimization cloud model

## Dynamic Graph Mining Method of the Minimum Energy Optimization Cloud

**Dynamic Picture.** The minimum energy optimization in dynamic cloud graph mining parallelism represents the first to solve the problem with a dynamic graph mining algorithms. Here are the concepts and definitions.

Definitions 1, dynamic map [5]. Let quintuple represents a dynamic map, which represents the set of vertices $G=(V,E,\Sigma,\lambda,P)$, $P$ represents a collection of edges for tag collection $\lambda:(V\cup E)\to\Sigma$, $t_e=(t\to t')$ mapping means that each vertex and each edge is assigned a mark for the weight set. In the period, the state transition diagram from one state to another, if it satisfies the conditions: (1) $V'=V$; (2) $E'=E-E_1$ or $E'=E\cup E_2$, where $E_1\subseteq E$, $E_2\subseteq(V\times V)-E$, $G$ is called Pictured dynamic map for the transition diagram $G'$.

Definition 2, transition probabilities dynamic graph [3] Given a dynamic map $G=(V,E)$ and a transition diagram $G'=(V',E')$, probability plots appear at moments into:

$$P\left((G\Rightarrow G')\middle|t\right)=\prod_{e\in E}P(e)\bullet\prod_{e\in E^{*}-E}\left(1-P(e)\right)$$

Equation $P(e)$ shows the transition probabilities side, there is a different side of this article assumes a dynamic figure $G$ is independent or not, based on probability and statistics shows that: For dynamic graph showing the available functions $P\left((G\Rightarrow G')\middle|t\right)$ in the sample $S(G)$ space probability distribution, which $S(G)=\{G_1',G_2',\ldots,G_n'\}$ represents FIG. collection.

**The Basic Idea of the Algorithm.** In a given period of time dynamic map database $GDB$, depth-first search algorithm, the search space for all child, first get frequent side and node, then expand the generation side of the candidate subgraphs after generation is complete, the calculation candidate diagram and dynamic support period $ES_{GDB}(g|t_e)$, if $ES_{GDB}(g|t_e)$ is less than the minimum support threshold, it is frequent, and continued all hypergraph depth-first search, if $ES_{GDB}(g|t_e)$ is less than the minimum support is not frequent, according to the previous definition, dynamic support apriori nature meet, so we can see all over showing the non-frequent, stop the second depth-first search, and return to the previous search space. Candidate subgraph generation process, $ES_{GDB}(g|t_e)$ taking into account the formula for calculating dynamic support $P\left((g\subseteq G)\middle|t_e\right)$, there is a probability in the calculation of the dynamic support, we need to be calculated, and for any two subgraphs $ES_{GDB}(g|t_e)\le ES_{GDB}(g|t_e)+ES_{GDB}(g_1|t_e)$, the right is the upper limit of the dynamic support subgraph, in the search space, the first to get support, so you can use prior knowledge to perform the cutting operation candidate subgraphs.

**Algorithm Design.** Optimized for minimum energy cloud model, select a computing node as the master node, control the operation of the entire program. The first phase of excavation edge set with a collection of nodes, the algorithm MEV (Mining Edge and Vertex) is described as follows:

Algorithm 2: MEV

Output: edge set collection $E\_List$ of nodes $V\_List$

Input: Dynamic Map Database $GDB$

Map Action: Scan dynamic map database $GDB$, the diagram format $<id,G>$, wherein the icon number, a string representation of FIG respectively count the number of nodes and edges and an intermediate sub-key pair $<e|v,num>$ and sent to Reduce this key operation .

Reduce: Accept the key value pairs and scanning operation, according to identify nodes and edges in ascending order, the same node identification and edges merge statistical side with nodes and edge weights were added to the side of the collection $E\_List$ and the node in the collection $V\_List$ .

Getting all sides and after the junction, get frequent side sets $FE\_List$ and frequent point $F1\_List$ set by a serial algorithm, randomly selected by the master node $Driver$ performs a computing nodes. GF1 algorithm process (Generate Frequent 1) perform the following:

## Experiment

**Experimental Environment and Experimental Data.** Experimental verification task are allocation algorithm, the minimum energy optimization cloud model and effectiveness of the dynamic graph mining and operational efficiency of the algorithm. Experiments in schools project to build the required hardware and software environment training center, where the hardware environment in the following table:

Table 2 Hardware configuration table

| Hardware | Hardware Description | Num |
|---|---|---|
| Rack Server | 2 × E5560 QC 2.8Ghz processor, 8GB PC3-8500DDR3 memory, 2 × 146GB 10K hard disk | 1 |
| Blade servers | High-performance HS22 blade servers, 2 × E5560 QC 2.8Hz processor, 8GB PC3-8500DDR3 memory, 2 × 146GB 10K hard disk | 10 |
| Fabric Switches | Center storage fabric switches containing 24 * 4Gb Shortwave SFP module port | 2 |
| Storage Array | DS5300 storage server, 28TB optical drive | 1 |
| switch | Cisco 6509,96 electrical port switch, dual redundant power supply | 1 |

The Taiwan rack servers as the master node, the other 10 blade servers as compute nodes, storage and fast access to data through optical switching and storage display, to meet the needs of massive data mining

**Experimental Design and Analysis.** Experiment is divided into two parts, the first part of the experiment on simulated data sets, in the calculation of the energy consumption platform, the need to monitor the operation of the system, record the number of computing nodes running elements, the number of task execution time in order to get the system idle energy consumption and execution, and finally get the average energy consumption.

The second part of the experiment is to test the minimum energy optimization cloud model and dynamic graph data mining algorithms in a real environment. The minimum energy optimization experiments with traditional cloud platform cloud model, the results as shown below:

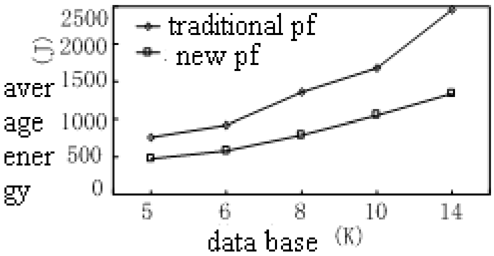

Fig. 2 Real environment average power consumption test results

## Conclusion

This paper presents an optimization based on minimum energy cloud model and large-scale dynamic graph mining algorithms to solve the massive problem of graph mining. Experimental results show that the algorithm is effective and feasible, with high efficiency, while a certain extent, reduce system power consumption. Future work will continue to optimize the minimum energy cloud model and dynamic graph mining algorithms to further improve mining efficiency and reduce system power consumption.

## Acknowledgements

Project: A Project Supported by Scientific Research Fund of Hunan Provincial Education Department

Project number: 15C1121

## References

[1] H.P. Tan, H.J. Wang, X.Y. Du. Dig Data Analytics -Rdbms with Mapreduce. Journal of Software, 2012,23 (1): 32-45.

[2] Tdwi Checklist Report: Big Data Analytics. Http://Tdwi.Org/Research/2010/08

Big-Data-Analytics.Aspx.

[3] Z. Zou, J. Li, H. Gao. Mining Frequent Subgraph Patterns from Uncertain Graph Data. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(9): 1203-1218.

[4] Potamias M, Bonchi F, Gionis A. K-Nearest Neighbors In Uncertain Graphs// Proceedings of the Very Large Data Bases. Singapore, 2010: 997-1008.

[5] Z.H. Jie, S.X. Jiang, Zou trillion in efficient top-k uncertain chart neighbor query processing algorithm. Journal of Computers, 2011,34 (10): 1885-1896.

[6] M. Hua, J. Pei. Probabilistic path queries in road networks: Traffic uncertainty aware path selection// Proceedings of the 13th International Conference on Extending Database Technology. Lausanne, Switzerland, 2010: 347-358.

[7] M.S. Li, Zou trillion years, Gao Hong, et expected calculate the shortest distance. on uncertain graph Computer Research and Development, 2012,49 (10): 2208-2220.

[8] W. Cai, B.L. Zhang, Lv Jianhua research surest maximum flow algorithms. uncertain graph Journal of Computers, 2012,35 (110): 2371-2380.