A dynamic statistical method based on peer grouping for spatial data access laws in a distributed system

PAN Shaoming, XU zhengquan, Liu Xiaojun
State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing
Wuhan University
Wuhan, Hubei, R.P.China
pansm@lmars.whu.edu.cn

***Abstract*—The spatial data access laws reflects the general characteristics of user preferences in tile access. According to the access laws, the strategy of storage and organization can be adjusted and utilized to store the spatial data in right place. However, the tile access has dynamic features(server peer capability, storage device and hot tiles have dynamic features). Dynamic statistical method are an important theoretical basis for improving the accuracy of storage and organization. A dynamic statistical method for the access law of spatial data based on the peer grouping is proposed in this paper to resolve above-mentioned problems. All distributed server peers are grouped and the size of group is controlled to reduce the flow rate of exchanging spatial data access information each other. The capabilities of the server nodes are calculated in this algorithm. And the leader peer with good service capabilities are chosen preferentially in the group to fuse dynamic statistical information. The experimental results show that this method has a shorter average fusion time for the dynamic statistics of spatial data access laws and thus can improve the efficiency of the large scale distributed spatial information service system.**

Keywords-Grouping; Spatial data; Distribution law; Dynamic statistics

I. INTRODUCTION

Along with the Internet technology's development and the popularity of broadband network. Spatial data service has changed from the desktop to the distributed network environment. The quality of spatial data server is very important to accelerate the promotion and popularization of spatial data service[1].

The popularization and application of distributed virtual geographic environment are still limited by a lot of practical conditions, such as the limited network bandwidth resources, the massive spatial data information and the limited storage space.

The strategy of storage and organization can be adjusted utilizing the access and distribution law of the spatial data. According the access and distribution law of the spatial data, the certain spatial data will be stored in a suitable storage medium and storage server peer, which will significantly improve system performance of spatial data services[2].

D Fisher proposed a Hotmap model based on the historical log information of the user behavior to request tiles[3], and the model considered that the access and distribution law of the spatial data consistent with a Power-law, thus there are small quantities of hot spot data which attract a large number of requests. Therefore, the hot spot data can be stored in High-Speed storage device to improve the access speed and the other spatial data can be stored in Low-Speed storage device to save cost.

However, the access and distribution law of the spatial data will change with the passage of time, but the Hotmap law is only a static model.

This article presents a dynamic statistical model in a distributed system. The model is based on peer grouping and choose the leader node to fuse dynamic statistical information. The size of group is controlled to reduce the flow rate of exchanging spatial data access information each other. This results in a shorter average fusion time for the dynamic statistics of spatial data access laws, thus improving the efficiency of the large scale distributed spatial information service system.

II. RELATED RESEARCH

There are many research results based on Hotmap Model.

According to the Hotmap model, J Krumm[4] then designed the system which can very accurately match the actual operation of the system status, and can greatly improve the performance of the system highly.

H. Wang proposed the access probability of a tile follows a Zipf distribution[5]. The Zipf law shows that the characteristics of users' access behavior follow a distribution law with parameter (α), as stated in Equation (1):

$$F_i = \theta / i^{\alpha} \quad (1)$$

Where F_i is the frequency of the *i*th most frequent tile, and θ is a normalized constant.

Zipf's law indicates that a tile which has frequently been accessed in the past may be requested again in the near future with a high probability. In other words, access to tiles has the property of 'Repeat Access'.

R. Li analyzed characteristics of users' access traffic for geospatial data (tiles), and based on Zipf law, constructed a mathematical model to simulate users' access data and studied the attributes of highly popular tile objects which have the great mass of access traffic[6].

III. THE DYNAMIC STATISTICAL MODEL

We can calculate that different statistical node scale produce different amount statistical information.

Published by Atlantis Press, Paris, France.

More statistical nodes will produce more efficiently statistical information and contain more widely statistical ranges, and so more benefit we will obtain from the statistical process. But at the same time, with the increasing of statistical object, the time will become longer, the result will be later and more invalid.

Conversely, less nodes will produce less efficiently statistical information (most of information are zero) and so less benefit we will obtain from the statistical process. But we can obtain the statistical information more quickly(Because of transmission faster). Therefore we must get the balance between the benefit and the real time. On the other hand, how to choose the leader node are very important in peer groups. The leader node is the key of transmission efficiency and transmission delay, because the node which have good service capabilities maybe attract more service request and its service load may be greater.

Based on the above reasons, we must consider not only the server peer capabilities, but also the size of group to solve the problem of dynamic statistics.

A. Algorithm Description

On the distributed virtual geographic environment, each server peer join a certain group according to its location and relationship with other peers.

Each peer has a neighbor peer list and send their own peer server capability information to other peers in the group at every scheduling period. This kind of information is called peer capability mapping(PCM).

According to the PCM information, the appropriate peers are chosen as the leader peer to provide agency services. At the same time, the size of group is controlled. This kind if algorithm is called GDSA (Grouping based Dynamic Statistical Algorithm), and that normal algorithm is called NDSA(Normal Dynamic Statistical Algorithm). The process of GDSA algorithm include several steps:

(1)Firstly, according to the strategy of grouping, all nodes are organized to several groups, and every group selects their own leader node, and then all leader nodes are organized to several leader groups, and so on.

(2)Secondly, when a group complete the data fusion, the leader node will process the information continue in the high-level.

(3)Lastly, when the highest-level node obtains the whole data fusion information, it will spreads the whole information to the low-level nodes through the reverse path.

The service node not only provide dynamic statistics service in the group, but also provide access service for remote user.

where S is the total number of server. If p_i is the ith node, b_i (Bps) is the total bandwidth of p_i(represent the peer capability), r_i is the total request queue length(represent the peer load) and u_i is the total terminal users of p_i(represent the peer potential load). b_i can be estimated by each node self[7], then the PCM information of p_i can be defined as: PCM(i)={ b_i, r_i, u_i }

B. Peer Selection Algorithm

The different peers in the group have different server capability, a certain peer will be chosen as the leader to fuse dynamic statistical information. The paper consider not only the peer capability, but also the peer load(include the potential load).

This is mainly based on the following three aspects:(1)The peer have larger bandwidth owns higher service ability, can provide more downloading service and faster downloading speed, but at the same time, the total request queue length will increase. And (2), the peer have shorter request queue length can response request faster. Then (3), the peer have more terminal users will receive more service request, the total request queue length and the network load will increase, thus the service ability will drop more quickly.

For a certain peer p_i, if F_i is the peer providing capability supported by p_i, and the peer have largest F_i will be selected to provide agent service. If there are two or more nodes which have the same F_i, we can choose it random from them. Then the F_i can be defined as Equation (2):

$$F_i = \frac{S \times b_i}{r_i \times \sum_{j=1}^{S} u_j + u_i \times \sum_{j=1}^{S} r_j} \qquad (2)$$

where b_i represent the peer capability. The peer which have larger bandwidth will be choose much more possible as leader. $r_i \times \sum_{j=1}^{S} u_j$ represent the peer load and $u_i \times \sum_{j=1}^{S} r_j$ represent the peer potential load. The peer which have larger load or potential load will be choose less possible as leader.

Equation (2) shows that, under the same conditions, the algorithm prefer to choose the peer which have larger service capacity, lower load or potential load to provide agent services.

C. Experiments and analysis

The experiments performed in this article used 90m of global Shuttle Radar Topography Mission (SRTM) terrain data for simulation.

The GDSA method was simulated for a distributed system, and the results were compared with NDSA method.

The average fusion time t gives an indication of the speed of statistical information fusion.

The first experiment simulate of 10-200 server peers and each peer load 150-250 terminal client users, and the bandwidth of peers obey [75Mbps, 125Mbps] uniform distribution and the average bandwidth of terminal client are 512Kbps. For GDSA method, all the peers are divided into two groups, and for NDSA algorithm, all the peers are in a group.

Figure 1 shows a comparison chart of the average fusion time of the two methods for several times and it's fitting curves. The fitting equation shows as Equation (3):

$$\begin{cases} t_{GDSA} = 1.0706 \times n + 3.31 \\ t_{NDSA} = 1.9216 \times n + 1.91 \end{cases} \quad (3)$$

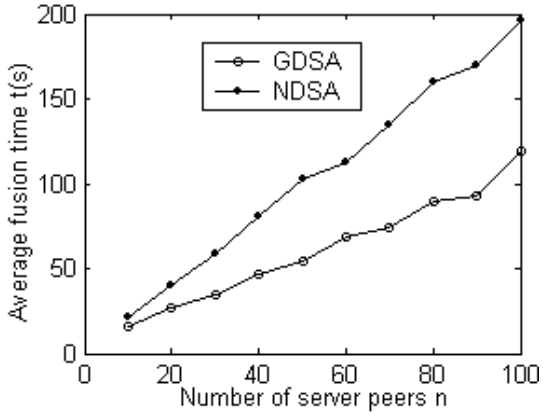

Figure. 1 The average time in different model

From Equation (3), the slope ratio between NDSA and GDSA is 1:0.56, that is to say, for NDSA, the GDSA have the slower average fusion times increasing speed with the peer scale expansion, and the speed is about 44% slower than NDSA. The experiments shows that the algorithm can meet the need of dynamic statistics in a large scale peer mode with high efficiency.

The second experiment simulate of 1000 server peers, all peers are divided into 1 to 200 groups(1, 2, 4, 5, 8, 10, 20, 25, 40, 50, 100, 125, 200).

Figure 2 shows a experiment chart of the average fusion time of GDSA methods in different grouping scale.

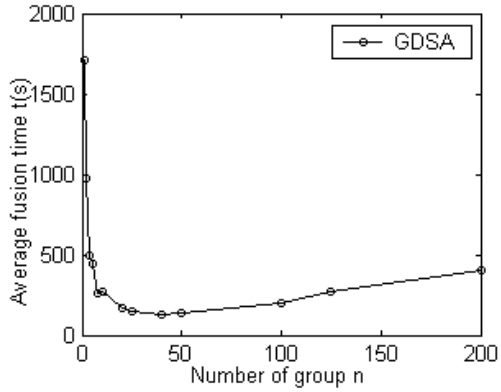

Figure.2 The average time in different grouping scale

From figure (2), when the number of group less than 20, the average fusion time will descend with the grouping scale expansion, but when the number of group more than 20, the average fusion time will increase with the grouping scale expansion.

The analysis shows that the load balancing algorithm can be used to average the differences of peers performance in distributed system when the peer scale is small. But with the grouping scale expansion, the leader peer load will increase and effect the performance.

IV. CONCLUSIONS

The spatial data access laws reflects the general characteristics of user preferences in tile access. Because the tile access has dynamic features, the access and distribution law of the spatial data is sorely in need of dynamic statistics. But the average fusion time will increase with the grouping scale expansion in large-scale distributed system. In this article, a dynamic statistical method for the access law of spatial data based on the peer grouping is proposed. In the method, all peers are divided into the several different groups, and the leader peer is chosen to fuse dynamic statistical information according their service capabilities.

The experimental results show that this method has a shorter average fusion time for the dynamic statistics of spatial data access laws and thus can improve the efficiency of the large scale distributed spatial information service system. Further study of statistical information compression algorithm in a distributed environment system can be expected to improve statistics accuracy.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. 41271398), the National Key Basic Research and Development Program of China (No. 2011CB302306), and the LIESMARS Special Research Funding.

REFERENCES

[1]. S Li, D Coleman. Modeling distributed GIS data production workflow.[J] Computers, Environment and Urban Systems, 2005, 29(4):401-424.

[2]. Yang C H, Wong D W, Yang R X, et al. Performance-improving Techniques in Web-based GIS. International Journal of Geographical Information Science, 2005,19(3):319–342.

[3]. D Fisher. Hotmap: Looking at geographic attention. IEEE Transactions on Visualization and Computer Graphics[J], 2007, 13(6):1184-1191.

[4]. J Krumm, E Horvitz. Predestination: Where Do You Want to Go Today? Computer[R], 2007, 40(4): 105-107.

[5]. H. Wang, S. M. Pan, M. Peng and R. Li, “Zipf-like Distribution and its Application Analysis for Image Data Tile Request in Digital Earth,” Geomatics and Information Science of Wuhan University, vol. 35, pp.356-359, 2010.

[6]. R. Li, W. Feng, Z. Xu, S. M. Pan. A mathematical simulation model for access traffic of geospatial data[C]. The 7th International Conference on Computer Science & Education (ICCSE 2012), Melbourne, Australia, 2012.

[7]. Zhang M, Xiong Y, Zhang Q, et al. On the Optimal Scheduling for Media streaming in Data-Driven Overlay Networks[J]. in: Proceedings of IEEE Global Telecommunications Conference (GLOBECOM’06), San Francisco, California, USA: IEEE Computer Society, November 27 -December 1, 2006, 1-5.