LETTER

An Association Rule Based Grid Resource Discovery Method

Yuan LIN[†a)], ***Nonmember***, **Siwei LUO**[†], ***Member***, **Guohao LU**[†], ***and*** **Zhe WANG**[†], ***Nonmembers***

SUMMARY There are a great amount of various resources described in many different ways for service oriented grid environment, while traditional grid resource discovery methods could not fit more complex future grid system. Therefore, this paper proposes a novel grid resource discovery method based on association rule hypergraph partitioning algorithm which analyzes user behavior in history transaction records to provide personality service for user. And this resource discovery method gives a new way to improve resource retrieval and management in grid research.

key words: *grid, grid resource, resource discovery, association rule, hypergraph partitioning*

1. Introduction

Grid, as a new type of distributed computing system, is a scalable, distributed infrastructure for sharing large number of heterogeneous resources in a distributed network for cooperative problem solving. It aims at coordinating resources more effectively in grid to provide various services for users. During the past few years, researchers have dedicated themselves to solve those problem in heterogeneous resource-sharing and to overcome limitation of computing power and storage capacity in grid [1]. As we all know, the primary service in grid is resource discovery that can return resource sets to match resources as description from users' resource requests [2]. Resource discovery mechanism usually store the information of resources based on the static grid system configuration in advance, and the users obtain the information of resources and find out resources suitable for their tasks through the mechanism. Therefore, a reasonable grid resource discovery method should provide effective resources management and organization mode which can reduce the size of candidate resource set, shorten the computing time of the resources matching and enhance the efficiency of resource scheduling.

The existing grid resource discovery methods come from traditional distributed computing system and search resources by heuristic algorithms which need traverse the whole resource nodes in system according to the resource requests of users. Most of the grid resource discovery methods are based on the simple low level resource characteristics and aim at computing scheduling, which mainly concentrate in the algorithm optimization of Min-Min, Max-Min, Sufferage and so on. There are also some methods such as flooding and random walk in P2P network [3], but these methods show limitation in flexible grid environment and are insufficient to meet the requirements for complex information environment and accuracy of resource retrieval. Meanwhile, since the grid resources are getting vast, scattered, and heterogeneous with Internet development, the resource discovery methods based on Web technology begin to show performance insufficient. The immediate sequel is returning so many resource retrieval results that scheduler cannot refine and classify these results to find the users' subjective needs. Currently, the ontology based resource discovery methods rises which abstract resource logic description from user's behavior model and resolve the problem brought by keyword-based information organization and retrieval [4].

Through analyzing those existing resource discovery methods in grid, we can deduce that resource in grid is lack of complete representation since resource description language is not unified in heterogeneous system and the process of resource discovery neglects user's behavior or interest while some researches show that the user's interaction behavior or interest is stable relatively [5], [6]. In order to make it possible to obtain what the users really need quickly and accurately from the resource query results, it becomes more meaningful to research on resource discovery method based on semantics and interaction behavior from resources scheduling processing. Therefore, we can not only consider the portability, scalability, and efficiency in service-oriented grid environment particularly, but also grid need to provide personalized services precisely. To achieve this goal, this paper proposes a novel resource discovery method which is based on association rule hypergraph partitioning algorithm (ARHP). This ARHP method analyzes user behavior from the history records of interactive information, and builds an associate grid resources database. Through clustering resources based on user's behavior and semantic from transaction records in grid, the ARHP method also shows a new way to grid resource management. The ARHP method has several advantages:

1) It can handle heterogeneous grid resources. Because grid is a heterogeneous system with large-scale resources in which resources characteristics description is different in different workspace, resource management become more and more complex in the open computation environment. But all of these problems can be minimized or solved by mining user behavior;

2) It establishes the correlation of resources and the grid sys-

Manuscript received November 25, 2010.

[†]The authors are with School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044 China.

a) E-mail: 06112070@bjtu.edu.cn

DOI: 10.1587/transinf.E94.D.913

tem can automatically adjust the relationship between resources according to the behavior migration of users;
3) The ARHP method orients users' behavior and can retrieve resources meeting users' needs, rather than those traditional resource discovery methods which search resources all nodes in system or simple keyword based;
4) This method develops from the original resource discovery methods, breaks through the traditional grid resources discovery based on resource description with simple and low level characteristics. It can be applied to semantics and mixed heterogeneous grid environment.

2. Grid Resource Discovery Based on Association Rule Hypergraph Partitioning Algorithm

Obviously, human is the grid end user. Therefore, it is very important to capture and model interactive behavior characteristics from user's transaction records. The ARHP based resource discovery method in this paper only need to collect user's behavior in grid instead of registering resource address and simple characteristics description through traversing all nodes in grid. The ARHP method analyzes and mines users' historic records to set up a resource transaction database, then associates all resources in grid by association rule to generate grid resources hypergraph which can be partitioned to clustering resources based on users' behaviors and semantics. This section will first give an overall idea of the model and then introduce the key steps of the ARHP resource discovery model.

2.1 Grid Resource Retrieval Based on Association Rule Hypergraph Partitioning Algorithm Model

The association rule hypergraph partitioning algorithm is a kind of data mining method to learn relation between information data and mainly focus on the relationship between structure and logic in data subset. As a result, it can handle more complex semantic relations and find higher level knowledge and the potential information. This ARHP method is more accurate, efficient and robust [7], [8]. The main steps in ARHP based resource discovery method are as follows:
1) To establish grid resource transaction database. When grid discoveries a node, it can collect resource scheduled records in the node and set up grid resources transaction database.
2) To get hyperedge through associating resources in grid based on hypergraph theory.
3) After obtaining the hypergraph, the hMETIS tools can partition the related resources into closely correlated resource set.

2.2 Grid Resources Association Rule Mining

In order to make grid resource discovery provide accurate resource retrieval service which can feedback resources fitting the users' real needs, the ARHP based resource discovery method will construct the correlation between grid resources. Let $< Q, R >$ be a transaction result which records resource set fed back as user's resource query, in which Q is the user's query description. Let R be $R = \{r_1, r_2, r_3, \ldots, r_m\}$ as scheduled resources set. According to the transaction results above, the system can set up a resource transaction database which is a kind of relationship table of resources in a row of which represents a resource query description consist of unique transaction ID and several accessed resources in a grid scheduling process. Every attribute in the table represents a resource in grid and the value of these attributes is 0 or 1, and the resource used in the scheduling process is labeled 1, otherwise 0. Then we set $TD = \{T_1, T_2, \ldots, T_l\}$ as resource transaction database and $T_i = \{r_a, r_b, r_c, \ldots, r_i\}$, $(r_a \in R, r_b, \ldots, r_i \in R)$ and $r_a \bigcap r_b \bigcap \ldots \bigcap r_i = \emptyset$ as a transaction record. All this information can be obtained in resources register process easily.

2.3 Hypergraph's Expression and Production

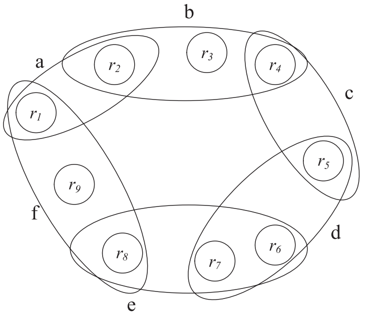
A hypergraph is a generalization of a graph and it is proposed by French mathematician F. Bouille in 1997. Hypergraph is composed of vertex and hyperedge. Usually we set $H = (V, E)$ as a hypergraph, the element $v_1, v_2, \ldots, v_m$ in vertex set $V = \{v_1, v_2, \ldots, v_m\}$ is the vertex in hypergraph and the element $e_1, e_2, \ldots, e_n$ in $E = \{e_1, e_2, \ldots, e_n\}$ is hyperedge of the hypergraph. There are a lot of differents between hypergraph and graph. Hyperedge in hypergraph can connect more than two vertices, which is said that a hyperedge can express the relationship between the numbers of vertices. Hypergraph modeling is an unceasing abstracted and reduced process in which original information wasn't lost in principle [8].

According to 2.1, we use the traditional Apriori algorithm in this paper to generate the hypergraph which is based on association rule [9]. This algorithm is the most influential association rules algorithm to mine frequent item sets in Boolean data. Apriori algorithm (shown as algorithm 1) can output resource hyperedge set E.

Algorithm 1. Association rules based Hypergraph generation

Input: transaction database TD
Output:the frequent item sets E in TD as hyperedge set of resource hypergraph

```
  do
    $min\_support$
      $= min\{support(r_1), support(r_2), \ldots, support(r_m)\}$,
      $r_1 \in R, r_2 \in R, \ldots, r_m \in R$
  $L1$= Large 1-itemsets;
  $For(k = 2; L_{k-1}! = 0; k + +)$ do begin
      $C_k = apriori - gen(L_{k-1})$;
        for each $T_i \in TD$ do
            $C_t = subset(C_k, t)$;
      for all candidates $c \in C_t$ do
            c.count++;
```



```
Hyper_graph File
6    9
r1   r2
r2   r3   r4
r4   r5
r5   r6   r7
r6   r7   r8
r1   r9   r8
```

Fig. 1 hMETIS format of grid resources based unweighted association rule hypergraph.

```
        end
          $L_k = \{c \in C_k | c.count \geq min\_support\}$
    end
$AnswerSet = \cup E_k$
```

2.4 Grid Resource Organization Based on Hypergraph Partitioning Algorithm

After obtaining resources association hypergraph in 2.3, we can partition the hypergraph into several resource clusters with closely relationship by hMETIS tools from university of Minnesota. hMETIS is a software package for partitioning large hypergraphs, and the algorithms in hMETIS are based on multilevel hypergraph partitioning which can successively reduce the size of the hypergraph as well as further refine the partition to produce high quality bisections. The computing process in hMETIS is very fast with fewer recursion times in large data set. The usage of the hMETIS can be obtain in hMetisManual [10]. The resource hypergraph shown in Fig. 1 has six unweighted hyperedges. The number of the vertices in the hypergraph is 9. So the resource hypergraph can be formatted as hMETIS input file (shown as Fig. 1).

3. Experimental Results and Analysis

The ARHP based resource discovery method mainly improves the precision of resource retrieval and it decreases resource search space in theory, especially in grid environment with massive resources. The experiment results below show that the ARHP based resource discovery method is effective. In the paper, we take the precision of resource retrieval as the proportion of scheduled resources to retrieved resources.

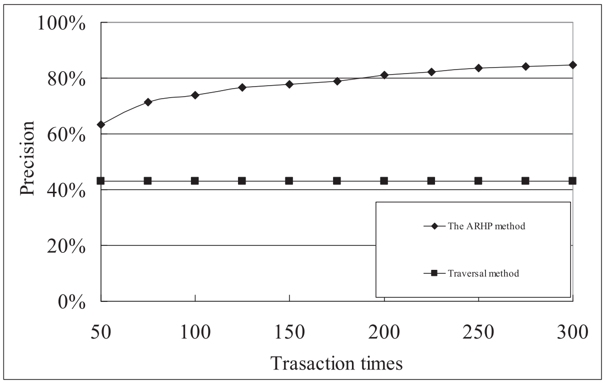


Fig. 2 Precision of resources retrieval in different transaction times.

Table 1 Results from the two data sets.

Method	Num. of Hyperedge	Num. of cluster	Precision
D1	13487	20	81.2%
D2	57963	20	83.7%

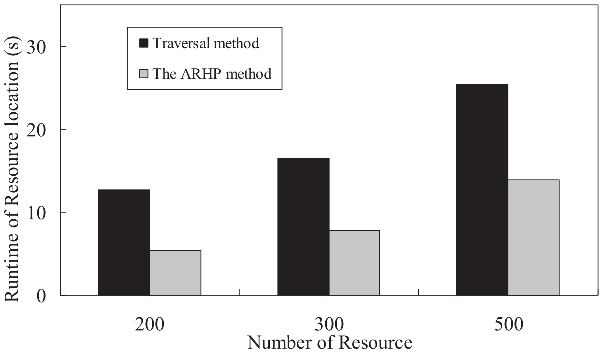


Fig. 3 Runtime of resource location with respect to different resource number.

Experiment 1 There are 500 resources in the experiment and the hypergraph is partitioned into 20 clusters. The result in Fig. 2 shows that the precision of resource retrieval in the method is improved gradually with transaction times increasing.

Experiment 2 According to different grid resources sizes, we contrast two dataset D1 and D2 and observe the results in the experiment. It shows that the precision of resources retrieval keeps relatively good and stable when resource scale changing. There are 500 resources, 200 transaction records in D1 and 1000 resources, 500 transaction records in D2. The results are shown in Table 1. From this experiment, we can find the ARHP method should work well in different system.

Experiment 3 Observing average runtime of random resource location 20 times with respect to 200 resources, 300 resources and 500 resources, the ARHP based resource discovery method is efficient compared with traditional traversal resource discovery method. The results shown in Fig. 3 suggest that the ARHP based resource discovery methods can reduce search space of resource to im-

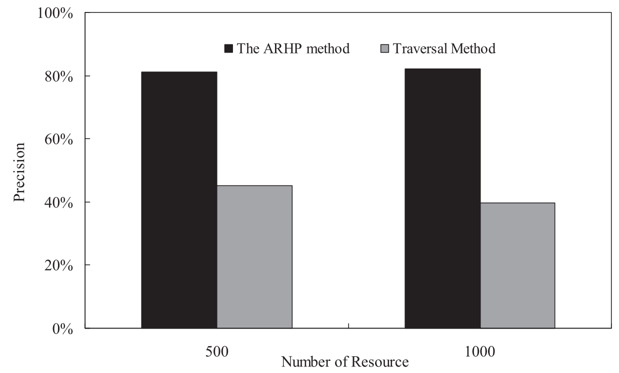


Fig. 4 Comparing precision of grid resource retrieval with traditional method.

proved location speed.

Experiment 4 For the two dataset D1 and D2, the ARHP based resource retrieval is efficient compared with traditional traversal methods. We can find the ARHP method can retain the precision of resource retrieval, while the traversal method the precision will reduce for the increasing in resources number, shown as Fig. 4.

4. Conclusion and Future Work

This paper presents a novel grid resource discovery method in the grid environment, which leverages user's behaviors to associate resources and improve resource retrieval and organization in grid. The ARHP based resource discovery method orients user and service and it reduce complexity of resource location. Especially, the method just need expand present grid system with low cost and scalability. However, the utilization of user behavior is critical in designing a system. In future, we should dedicate ourselves to research on user's behavior in grid.

Acknowledgements

This work is supported by National High-Tech Research and Development Plan (2006AA01A121) and Development Grant for Computer Application Technology Subject jointly Sponsored by Beijing Municipal Commission of Education and Beijing Jiaotong University (XK100040519).

References

[1] I. Foster and C. Kesselman, Grid2, Blueprint for A New Computing Infrastructure, San Francisco, Morgan Kaufmann Publishers, 2004.

[2] A. Iamnitchi and I. Foster, "On fully decentralized resource discovery in grid environments," Proc. Second International Workshop on Grid Computing, pp.51–62, London, UK, 2001.

[3] Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker, "Search and replication in unstructured peer-to-peer networks," Proc. 16th ACM Int. Conf. on Supercomputing. pp.84–95, ACM, New York, 2002.

[4] W.C. Shih, C.T. Yang, and S.S. Tseng, "Ontology-based content organization and retrieval for SCORM-compliant teaching materials in data grids," Future Generation Computer Systems, vol.25, no.2009, pp.687–694, 2009.

[5] A. Asvanund, R. Krishnan, M. Smith, and R. Telan, "Interest-Based self-organizing peer-to-peer networks: A club economics approach," Proc. 13th Workshop on Information Technology and Systems, http://heinz.cmu.edu/research/175full.pdf, 2004.

[6] A. Crespo and H. Garcia-Molina, "Semantic overlay networks for P2P system," Proc. 3rd Int Workshop on Agents and Peer-to-Peer Computing, pp.1–13, Springer, Berlin, 2005.

[7] E.H. Han, G. Karypis, V. Kumar, and B. Mobasher, "Clustering based on association rule hypergraphs," Workshop on Research Issues on Data Mining and Knowledge Discovery, pp.9–13, 1997.

[8] E.H. Han, G. Karypis, V. Kumar, and B. Mobasher, "Hypergraph based clustering in high-dimensional data sets: A summary of results," Bulletin of the Technical Committee on Data Engineering, vol.21, no.1, pp.15–22, 1998.

[9] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," Proc. 20th International Conference on Very Large Data Bases, pp.478–499, Santiago, Chile, 1994.

[10] hMETIS A Hypergraph Partitioning Package.
http://glaros.dtc.umn.edu/gkhome/fetch/sw/hmetis/manual.pdf