Research on Service Usage Pattern Mining Method in the Distributed Context

Zhu Hong-Kang
School of Computer Science& Software Engineering
TaiYuan University of Technology
TaiYuan 030024, China
zhuhkxyx@126.com

Yu Xue-Li
School of Computer Science& Software Engineering
TaiYuan University of Technology
TaiYuan 030024, China
Xueli13287@263.net

Abstract: Composite service selection is a research difficulties in service-oriented computing domain. Most selection approaches are based on service QoS information which is difficult to acquire accurately. In this paper, a distributed web service usage pattern mining infrastructure based on services execution information is proposed. In order to meet the cross-organizational associated business requirement, a service registry federation mechanism is presented. Then, a distributed algorithm of composite service frequent sequential mining within registry federation is described. In the end, some simulation results are demonstrated to show the effectiveness of the proposed method.

Keywords: composite services; association rule; sequence pattern; data mining

I. Preface

The world has made a quite great progress in Web Service technology in recent years, which is used as the key way to implement service system infrastructure. Enterprises use Web Services to encapsulate business process and make this registered process freely accessible to their partners via Internet. In addition, single service may be dynamically bound up in Web Services to provide some new value-added composite service. This new pattern of software development successfully gets rid of repetitious development of service-oriented software products while materializing advantages of Web Service's support to quick application integration. However, substantial autonomy and heterogeneity of Web Services leads to a fundamental question during the implementation of composite service strategy: how can the strategy be ensured to be exactly and orderly executed?

Now we may easily record runtime tracks of composite service along with flood operation of such service, which objectively reflect operational conditions of composite service. Mining of aforesaid useful information may help us make better analysis, selection, monitoring, optimization and improvement of composite model. In fact, we must take account of the below actual effects during recording and mining of service execution information in the context of real service computation:

1) Since enormous Web services are widely distributed into Internet and organized by different institutions, distributed Service Registration Federation is hence required to maintain description of such service and relevant service log database for recording execution information.

2) Cross-organizational service business association requires for different types of relations among distributed registration centers, for which a reasonable mechanism is necessary to build Service Registration Federation.

3) Distributed data mining technology is required to analyze service log database in Service Registration Federation.

This article is designed to provide P2P-based DSEMI (Distributed Service Execute Mining Infrastructure), probe into implementation mechanism of such alliance based on DSEMI, and provide specific log database-based distributed Web Service association rule mining algorithm in this alliance. They will be discussed in section 2, 3 and 4 respectively. Experimental analysis of mining algorithm will be introduced in section 5. Conclusion and expectation will be represented at last.

II. P2P-based DSEMI

To provide free support for multiple service registries and convenient access to recording and mining of service logs,
P2P-based (Peer-to-peer) DSEMI is then suggested in this article as shown in Graph 1.

DSEMI is consisted of the following three types of Peer: RM-Peer (Registries Management Peer), SL-Peer (Services Log Peer) and C-Peer (Client Peer).

RM-Peer is responsible for management of the whole infrastructure, such as creation of SL-Peer for registered service registries, establishment and maintenance of Service Registry Federation. It has to create SL-Peer for every Service Registry Center to record and mine execution information, whose functions may be seen like half section in the right of Graph 1: SOAP Intermediaries collect information necessary for service execution that will be kept in the log database in XML after selection and filtering; service mining tools use log database to support selection and ensemble of service. C-Peer is responsible for collect query information from customers and return query results of SL-Peer.

DSR (Distributed Services Registries) is one kind of service registration center interspersed among Internet. Composition of every R-R database (Registries-Repository) is shown in half section of the left of Graph 1. Every R-R database also provides service entity information base besides registration center issuing service registration information, such as WSDL document library, OWS-S document library and BPEL document library.

III. Establishment of Service Registry Federation

There is a strong association between different service registration centers interspersed among cross-organizational and cross-fields business integration. Therefore, it will be helpful to implement cross-fields service contracts by logically dividing these centers into different RFs and reduce mining space and increase mining accuracy by mining service execution information in different RFs. We then refer to References [2] to define RO (Registry Ontology) as shown in Graph 2. RO is established and controlled by RM-Peer and RM-Peer will create SL-Peer for a new Registry Center in DSEMI and issue updated RO to other SL-Peers in the Federation.

IV. Services Associate Rules Mining based on Registry Federation (SARM-FC)

Abundant execution information of R-R base is recorded in log database of relevant SL-Peer in the Federation, which also consists of interaction information of different centers. Therefore it may greatly increase efficiency of service selection and ensemble by effectively mining such information. This section is designed to study the way of providing mining methods of association rules in relevant distributed log database in the Federation.

A. Definition of Record of Log Database

We have to record various information provided by Definition 1 for each piece of interaction information among SOAP Intermediaries for the purpose of obtaining log database used by mining process. We suppose the same recording format is used in log database of SL-Peer for simplifying association mining algorithm.

Definition 1: XML Record is one data recorded by single service execution information in log database and may be expressed by six unit groups (Composite Services ID, Instance ID, Services ID, Type, Time Stamp and Status).

Composite Services ID: This composite service ID may be those BPEL document-based modes representing an abstract process of composite service.

Instance ID: This composite service execution instance ID is a unique execution example of identifying composite service.
Services ID: This service instance corresponding to abstract service in composite service may be uniquely identified by URI of example service.

Type: Type of SOAP information represents modes of request or response.

Time Stamp: It represents current moment of executing Services ID service.

Status: It represents the condition of success or failure of service request or response.

B. SARM-RF

Suppose there are $n$ SLPeers in the Federation with relevant log databases such as $\{DB_1, DB_2, \ldots, DB_n\}$. Composite Service Execution Instance Set hereinto may be viewed as the Transaction Set of association mining process for every $DB_i$ (every different execution instance will be uniquely labeled by Instance ID); suppose data set is $I_i=\{WS_{i,j}\in DB_{i,j}=1,2,\ldots,m\}$ and $WS_{i,j}$ represents the specific service corresponding to $DB_i$ Transaction Set. Suppose $DB=\bigcup_{i=1}^{n} DB_i$, then SARM-RF is designed to find Frequent Set of the overall $DB$ in $DB_i$ Log Database of Peer to Peer.

Apriori\(^3\), even as the classical algorithm for associate rule mining, is only limited to centralized data set. The key problem of associate rule mining in distributed data set is how to reduce information transmission quantity among different nodes. Some classical algorithms are available, including CD\(^4\), FDM\(^5\) and D-Sampling\(^6\). We herein select FDM as the prototype to construct SARM-RF algorithm in the context of Peer to Peer. Define remarks shown in Table 1.

The core idea of SARM-RF algorithm is to create local candidate set for each node, which will then create the overall candidate set according Theorem 1. Two types of pruning technologies are used to reduce information transmission between nodes, including Local Candidate Set Pruning and Overall Candidate Set Pruning.

Theorem 1 \(^5\): $\forall k>1$, the following formula is true:

$$L(k) \subseteq CG(k) = \bigcup_{i=1}^{n} CG_{i(k)} = \bigcup_{i=1}^{n} Apriori \_gen(GL_{i(i-1)})$$

Suppose $D = \sum_{i=1}^{n} D_i$.

<table>
<thead>
<tr>
<th>Remarks</th>
<th>explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_i$</td>
<td>number of transactions in $DB_i$</td>
</tr>
<tr>
<td>$D$</td>
<td>$\sum_{i=1}^{n} D_i$</td>
</tr>
<tr>
<td>$\minsup$</td>
<td>support threshold minsup</td>
</tr>
<tr>
<td>$GL_{i(k)}$</td>
<td>globally large $k$ itemsets</td>
</tr>
<tr>
<td>$CG_{i(k)}$</td>
<td>global support count of $X$</td>
</tr>
<tr>
<td>$CA_{i(k)}$</td>
<td>candidate sets generated from $L_{i(k)}$</td>
</tr>
<tr>
<td>$GL_{i(k)}$</td>
<td>$CA_{i(k)}=Apriori_gen(L_{i(k)})$</td>
</tr>
<tr>
<td>$LL_{i(k)}$</td>
<td>Gl-large $k$-itemsets at $S_i$</td>
</tr>
<tr>
<td>$LL_{i(k)}$</td>
<td>candidate sets generated from $CG_{i(k-1)}$</td>
</tr>
<tr>
<td>$X_Sup_{i}$</td>
<td>locally large $k$-itemsets in $CG_{i(k)}$</td>
</tr>
<tr>
<td>$X_Sup_{i}$</td>
<td>local support count of $X$ at $S_i$</td>
</tr>
</tbody>
</table>

SARM-RF algorithm:

Input: Transaction Database $DB_i$, support threshold $\minsup$

Output: $L$: the set of all globally large itemsets

Method: Iteratively execute the following program fragment distributively at each Peer. The algorithm terminates when either $L_{i(k)}=\emptyset$ or the set of candidate sets $CG_{i(k)}=\emptyset$

Begin

/* generate candidate sets */

If $k=1$ then

$$T_{i(1)}=get\_local\_count(DB_i, \emptyset, 1)$$

Else {

$$CG_{i(k)} = \bigcup_{i=1}^{n} CG_{i(k)} = \bigcup_{i=1}^{n} Apriori\_gen(GL_{i(i-1)})$$

$$T_{i(k)}=get\_local\_count(DB_i, CG_{i(k)}, \emptyset)$$

} /* Local Candidate Set Pruning and Overall Candidate Set Pruning */
for all \( X \in T_{i(k)} \) do
  if \( X.\text{Sup}_i \geq S^*D_i \) then
    if \( X.\text{Sup}_i + \sum_{j=1,j\neq i}^{n} \text{MaxSup}_j(X) \geq S^*D \)
      then insert \( <X,X.\text{Sup}_i> \) into \( L_{i(k)} \)
      /* broadcast \( L_{i(k)} \), compute gl-k itemsets */
      for \( j=1 \) to \( n \) do send \( L_{i(k)} \) to Peer \( S_j \);
    Receive \( L_{i(k)} \) from other Peer;
    For all \( X \in L_{i(k)} \) do {
      \( X.\text{Sup} = \sum_{i=1}^{n} X.\text{Sup}_i \);  
      if \( X.\text{Sup} > S^*D \) then 
        insert \( X \) into \( G_{i(k)} \);
    } /* Broadcasting results */
    broadcast \( G_{i(k)} \)
    receive \( G_{i(k)} \) from all other Peer \( S_j (j \neq i) \)
  \( L_{(k)} = \bigcup_{i=1}^{n} G_{i(k)} \);  
  Divide \( L_{(k)} \) into \( GL_{i(k)} \);  
  Return \( L_{(k)} \)
End.

V. Analysis of Simulation Experiment

This section will evaluate the efficiency of SARM-RF algorithm through simulation experiment. The experiment is conducted based on the prototype system of DSEMI, in which each Peer will be provided with one computer (Pentium(R) 4.3GHz and 1GB RAM). All Peers will be linked in 100Mb LAN.

Simulation experiment of Group 1 is designed to verify the effect of number of nodes in the Federation on performance of the algorithm. There are six different nodes in the experiment, each of which will randomly create 10000 pieces of instance data in 40 composite services. Service interleaving access ratio among nodes is 10% (Service interleaving access ratio = Number of service of interleaving access among nodes/Total of access service among nodes). Operation efficiency of SARM-RF is shown in Graph 3 in the context of maximum sequence length (7) and differently minimum support degree.

Simulation experiment of Group 2 is designed to verify the effect of service interleaving access ration on performance of the algorithm. There are still four different nodes in the experiment, each of which will randomly create 10000 pieces of instance data in 40 composite services. Service interleaving access ratios among nodes step from 5% to 30% (this value increases by degree of 5%). Operation efficiency of SARM-RF is shown in Graph 4 in the context of maximum sequence length (7) and differently minimum support degree.

The experiment results rudimentarily show operation efficiency of SARM-RF will be applicable within the acceptable linear scope along with the increasing of number of nodes of the Federation and interleaving access ratio. Currently it is no way to make further comparison due to scarcity of standard testing platform and cases or similar distributed service mining algorithm in the data search. Therefore experimental test just rudimentarily shows the logicality of DSEMI model and DSEMI-based SARM-FR.

VI. Conclusion and Expectation

This article innovatively suggests P2P-based DSEMI through service execution information as for selection of composite service, designs a unique mechanism to build Service Registry Federation for satisfying requirements on cross-organizational business association, and designs a
distributed algorithm to implement composite service frequent sequence mining among the Federation. Experimental design and analysis are also provided based on the given algorithm, showing availability and effectiveness of DSEMI and algorithm.

SARM-RF is hereinto used to perform data mining in the frequent mode, making FDM as the prototype. Comparing to other distributed sequence mining algorithms particularly application in huge service information mining, the effect of such algorithm still needs to be further verified and studied.

In addition, we seriously give priority in the future research on the subject of how to integrate service execution information based composite service mining method suggested by this article into traditional modeling method of composite service process for improving quality of selected composite service.

References:


