Based on Digital Archives for Scientific Workflow System Active Service

Rui Wang and Dailu Tang

ABSTRACT

This paper describes the architecture of active service model of scientific workflow system. Digital archives technologies are available for all the resources needed to construct services specific group. This paper reviews the introduction and motivation for active service approach, describes the architecture of scientific workflow system, discusses the technologies used in active service, which uses digital archives technology to enable multi-service organizations achieve scientific workflow goals.

INTRODUCTION

Workflow management systems (WfMS) provide the automation of science processes where a collection of tasks is organized between participants according to a set of defined rules to accomplish some science goals [1-2]. Often, traditional WfMS can only coordinate workflows and their enacting agents within a single organization. However, advanced WfMS can now interact with various types of distributed agents over the Internet [3]. A science process can be viewed as a process to solve a scientific problem with goals such as knowledge discovery, knowledge innovation, and so on. A large-scale complex science process needs to handle very complicated logics in scientific problems.
Although many research projects are exploring the problems about computation- and data-rich scientific collaboration, there have been very few attempts to investigate the architectural issues in scientific workflow system. Not only the number of organizations in the workflow cannot be easily predicted, but the available services may change over time, too. Moreover, it is desirable that services are provided in a user adapted and context-adapted way [4]. Science processes are more data-centric and knowledge-intensive and, active services will provide customized services to better satisfy the needs of every individual participants. Active services also enable science processes have highly creative, innovative and dynamic.

We present architecture for a scientific workflow system management offering active services, which uses digital archives technology to enable multi-service organizations achieve scientific workflow goals.

**SCIENTIFIC WORKFLOW SYSTEM**

A scientific workflow system contains a specification of a science goal. For a given condition regime, the goal set G will comprise the set \( \{g_0, g_1, \ldots, g_n\} \), where each \( g_i \) will have a relationship to one or more other \( g_j \). A part of certain workflow goal is implemented by one migrating agent regarded as a service consumer. Attributes include Goal_id, a set condition on goal achievement, and a validity time and cost.

Alliance develops an architecture that facilitates fault tolerant cooperative control of heterogeneous organizations towards a common workflow goal. By negotiation between organization partners, we mean all the conversation acts made between a requester partner need a workflow service and one or several selected provider partner(s) able to provide the requested service. Alliance is robust, reliable and flexible in dynamic and unpredictable environments. Attributes include Alliance_id, a set of organization, lifecycle.

Science organizations offer services to the workflow agents who visit them. Each organization has a service agent as a leader and may provide one or more services. Every service is a representation of a goal type that can be handled by this organization. Attributes include organization name, one anchorage (the place of migrating agents moving to), one or more workstations.
A specific group is a set of service agents that also determines a virtual interaction space: a service agent may communicate with another service agent only if they belong to the same group. The specific group gives prominence to the core competitive and cooperative ability of each member organization, and optimizing the value chains of group, then mapping a collection of comparable or related workflow services to a single logical service.

A workflow service is a computational entity which is able to achieve a service requestor goal, having different criteria (due time, price, visibility of the service evolution and way of executing the service). Our active service model is based on OWL-S, which is a set of ontology definitions for describing service behaviors. The active service agent is just another site, but provides only one service: s_info, which returns a list of service descriptions that match (exactly) the service name. On the one hand, the active service agent silently records histories of migrating agents’ interaction and builds profiles by observing their behavior. On the other hand, the active service agent aggregates and manages multiple service organizations as a single virtual service.

**Active Service Model**

Active service mechanism provides a distributed environment where multiple application organizations can coexist in virtual or physical resources, such that migrating agents are unaware of the complexities inherent to workflow-computing. The active service model makes extensive use of virtualization technology for the creation of dynamic specific groups of virtual services that can be aggregated on-demand.

Active service is based on matching abstracted workflow goal descriptions with semantic annotations of workflow services. This process can only happen on an ontological level. In a distributed environment, different users and web services may use different terminologies, which lead to the need for mediation in order to allow heterogeneous parties to communicate. Given the previous assumption, we can optimistically assume that a mapping has already been established between the used terminologies. For this, the concrete migrating agent need has to be generalized to more abstract goal descriptions, and concrete services and their descriptions have to be abstracted to the classes of services that the active service agent can provide.

**Acquiring Goals**

The first task of the active service agent is to identify data consumers’ needs and wants accurately, then to provide information of workflow organizations that will satisfy them. For forecasting to be successful, it is not sufficient to merely discover what service customers require, but to find out why it is required. Only by gaining a deep and comprehensive understanding of clients’ state and workflow context can active service’s goals be achieved. Therefore, active service agent attempt to visit the workflow engine which is the core of workflow management system to acquire migrating agents desires.

**Building Specific group**

A service specific group maps a collection of comparable or related services to a single logical service. It is managed by the active service agent consisting of a service entry component, a rule-based policy component, a service catalog, and a service-rendering component. Building service specific group’s objective is not to define new
application programming interfaces (APIs) or new standards, but to construct from the existing organizations a new, higher-level structure that can hide complexities from service consumers, simplify deployment for service supplies.

Let \( SWI = (O, E) \) be an information system of the science workflow system, where \( O \) is a non-empty finite set of organizations and \( E \) is a non-empty finite set of effect such that \( e: O \rightarrow S_e \) for every \( e \in E \), \( S_e \) is the service set for effect \( e \). With any \( EF \subseteq E \), there is an associated similar relation:

\[
\text{SIM}(EF) = \{(x, y) : x \in O, \exists e \in EF, s(x, e) = s(y, e) \} \quad \text{or} \quad \exists \theta, s(x, e) = s(y, e)
\]

If \( (x, y) \in \text{SIM}(EF) \), \( x \) and \( y \) are similar by effect from \( EF \). The similar class of the EF-similarity relation is denoted \( \text{SG} \text{(Specific Group)} \). \( O \text{SIM}(EF) \) defines a service group granule, then this is not partition but overlay of \( O \).

**Organization Using Digital Archives**

Three new digital archives and a selection mechanism are devised for organizations of specific groups. Competing operator and cooperating operator realize competition and cooperation among organizations. Self-learning operator increase the energy of organizations by knowledge.

On the basis of effect significance, matching degree of organization is calculated according to formula:

Given that \( o \in \text{organization} \) and \( r \in \text{rule} \), \( |\text{condition}| \) denotes the number of \( <\text{condition}> \) in the part of \( r \), and \( |\text{condition}^o| \) denotes the number of condition satisfied by organization \( o \).

The matching degree value:

\[
MD^r_o = \frac{|\text{condition}^o|}{|\text{condition}|}
\]

According to the definition, the range of the match value is \([0,1]\), while 0 is the worst case and 1 the best case. The rule with the maximum match value is used to predict the optimal organizations in this specific group.

- **Competing digital archives**

Performing competing operator on organizations in one specific group is to find the max matching degree of service organization. If \( \text{org}_i \) satisfy \( MD(\text{org}_i) = \max \{ MD(\text{org}^\text{max}) \} \), \( \text{org}_i \) is recorded as winner, and \( MD(\text{org}^\text{max}) = MD(\text{org}_i) \).

- **Cooperating digital archives**

Given that two parent organizations, \( \text{org}_i = \{ p_1, p_2, \ldots, p_n \} \) and \( \text{org}_v = \{ q_1, q_2, \ldots, q_n \} \). If they do not achieve goal respectively, then one child \( \text{org}^v \) is determined by cooperating strategy to expand influence and enforce energy:

- **Self-leaning digital archives**

All organizations in one service group could not satisfy the goal of conditions. Organizations improve their matching degree through self-learning from other alliance’s group to.

**PERFORMANCE AVALUATION**

To evaluate service recommendation based on participation matrix contract net (SRPMCN), we use F1 which is a well-known performance measure for recommender
system [5-6]. We start by dividing our data sets into two parts: the training set and the test set. SRPMCN works on the training set first, and then generates a set of recommended services, called service recommendation set. $F_1$ integrates recall and precision and is given by:

$$F_1 = \frac{2rp}{r+p}$$

where $r$ is recall, counting the ratio of the number of services correctly assigned to the test set $t$ to the total number of services belonging to the test set $t$; and $p$ is precision, counting the percentage that services assigned to the test set $t$ actually belong to the group service recommendation set $g$. For recall and precision, there exists contradiction between them. Emphasizing one side alone will lead to unacceptable low value in the other side. So $F_1$ metric is widely used as an optimization criterion for binary decisions. For a good service recommendation, the $F_1$ value should be high.

The population size has a significant impact on the recommendation quality. To determine the sensitivity of population size, we performed an experiment where we varied the number of population members’ from 10 to 30 and computed the corresponding $F_1$ metric. Our experimental results are shown in Fig. 3. Looking into the results, we can conclude that the size of the population members does affect the quality of service recommendation. Generally, the recommendation quality increases as the number of population members increases. However, after a certain peak, the improvement gains diminish and the quality becomes worse. In our experiment, the peak was reached at the area between 10 and 15.

![Figure 2. Impact of population member size on recommendation quality.](image)

**CONCLUSION**

This paper motivates the need for an architectural framework of active services in a heterogeneous workflow environment. In this paper, we have proposed a goal-oriented scientific workflow system based on the mobile agent technology. We introduced the concept of service specific group which mapping a collection of comparable or related workflow services to a single logical service. We discuss a way to do active service by allowing runtime selection, integration and coordination of distributed resources and also accommodate dynamic science requirements.
REFERENCES