Organizing Dynamic Multi-level Workflows on Multi-layer Grids for e-Business Portals Development

Jia Hu and Ning Zhong
Dept. of Systems and Information Eng., Maebashi Institute of Technology
460-1 Kamisadori-Cho, Maebashi-City 371-0816, Japan
hujia@kis-lab.com, zhong@maebashi-it.ac.jp

Abstract

Developing an intelligent e-Business portal needs various WI technologies. In order to represent, manage, integrate, share, and use the information coming from huge, distributed, multiple data sources, we present a conceptual model with dynamic multi-level workflows corresponding to a multi-layer Grid architecture for multi-aspect analysis in building an e-Business portal on the Wisdom Web, and for dynamically organizing status-based business processes that govern enterprise application integration.

1. Introduction

Web Intelligence (WI) is a new and active research filed derived from Artificial Intelligence (AI) and Information Technology (IT) [16, 17]. Developing intelligent e-Business portals is one of the most sophisticated applications of WI, which needs to be supported by various WI technologies [13, 18].

An e-Business portal enables an enterprise or a company to create a virtual organization (or a virtual community) on the Web where key production/information steps are outsourced to partners and suppliers.

In other words, an e-Business portal is a single gateway to personalized information needed to enable informed business decisions, in which all of the contents related to the virtual organization can be accessed although such organization information is geographically distributed in multisite, multi-data repositories, and multi-institution.

In general, there are two typical types of enterprises,

(1) transnational corporations that have operations, subsidiaries, investments, or branches worldwide,

(2) communities with many mid-size/small-scale companies in a region,

which need such portals for supporting their e-Business activities.

As the huge and multiple data sources are coupled with and geographic distribution of data, users, systems, resources, and services in the typical types of enterprises, Grid platform is an ideal middleware or platform for e-Business portals development. Grid concepts and technologies are originally developed to enable resource sharing within scientific collaborations, which provide mechanisms for sharing and coordinating diverse resources and thus enable the creation, from geographically and organizationally distributed components, of virtual computing systems that are sufficiently integrated to deliver desired qualities of service [7].

Furthermore, workflow management systems address the enterprise process automation problems, which refer to a formal, executable description of a business process [1]. A workflow management system is a software platform that supports the design, development, execution, and analysis of workflow processes.

In order to represent, manage, integrate, share, and use the information coming from huge, distributed, multiple data sources, we present a conceptual model with three levels of dynamic workflows, namely data-flow, mining-flow, and knowledge-flow, corresponding to three layers of Grids called data-grid, mining-grid, and knowledge-grid, respectively, for deploying and managing the data mining agents on the Grid for multi-aspect analysis in distributed, multiple data sources, and for dynamically organizing status-based business processes.

The rest of the paper is organized as follows. Section 2 discusses the background and motivation of our work. Section 3 gives the architecture of an e-Business portal and its main features. In Sections 4 to 6, we describe three-level workflows versus three-layer Grids in detail, respectively. Finally, we conclude the paper in Section 7.
2. Related Work

Grid computing attempts to provide a powerful middleware for coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations [8]. Although most of Grid projects have focused on resource sharing in the distributed environment, researchers are beginning to touch about how to employ knowledge processing on the Grid [4, 11].

How to develop an infrastructure focused on the design and implementation of an environment for geographically distributed high-performance knowledge discovery applications is discussed in [3], however, how to combine data mining and knowledge discovery with reasoning and using multiple information sources needs to be investigated at a unified way in depth.

Increasingly, it is becoming necessary to develop higher level services that can automate the process and provide an adequate level of performance and reliability. Meanwhile, E. Deelman et al discussed issues associated with workflow management in the Grid in general and provided a description of how to generate executable workflows on the Grid accordingly [5]. Furthermore, M. Mika et al discussed the problem of scheduling workflow applications on a Grid, which consists in assigning Grid resources to tasks of a workflow job across multiple administrative domains in such a way that minimizes the execution time of a particular set of tasks [10]. In [22], H. Zhuge proposed an agent-based cognitive workflow system for distributed team cooperation and knowledge management. Y. Gil et al used artificial intelligence planning techniques to compose valid end-to-end workflows on the Grid [9].

So far, most of the workflow systems on the Grid maintain a static specification with a single layer model [5], in this way they will be hard to tackle real world problems. The paper attempts to present a dynamic three-level workflow management system with respect to three-layer Grids for enterprise application integration.

3. The Architecture of an e-Business Portal

Figure 1 shows the architecture of an e-Business portal that has been developing by us. In this e-Business portal, there are mainly three kinds of data sources on the data-grid, namely customers DB, products DB, and Web-logs DB. Various data mining methods (i.e., targeted marketing, ordering rule, and Web-log mining) are employed as agents on the mining-grid for various service-oriented, multi-aspect data analysis [19]. Furthermore, knowledge mined from multiple data sources is employed to provide personalized services for customers, portal updating, and enterprise marketers through the knowledge-grid.

4. Knowledge-flow vs. Knowledge-grid

The three levels of dynamic workflows, namely data-flow, mining-flow, and knowledge-flow, corresponding to three layers of Grids called data-grid, mining-grid, and knowledge-grid, respectively, are generated dynamically, based on the conditions (situations), data quality analysis, and applications, to represent, manage, integrate, share, and use the information on the three-layer Grids.

In the multi-layer architecture, lower levels provide middleware support for higher level applications and services, thereby opening the door to developing more complex, flexible, and effective systems.

The architecture is defined on the top of Grid middleware and services, i.e., it uses basic Grid services to build specific knowledge services. Following the integrated Grid architecture approach, these services can be developed in different ways using the available Grid toolkits and services. The current implementation is based on the Globus toolkit and Tomcat, so that the Grid-enabled applications can be delivered to end users via a standard Web browser.

Figure 1. The architecture of an e-Business portal
4.1 Customer-driven Flow

A customer-driven flow is invoked by a series of customer activities, for instance, visiting portal, making registration, searching, browsing, shopping, and paying. In order to describe the workflow process, directed graph (or flow graph) is used, which defines the order of execution among the nodes in the process [1].

Figure 2 shows modeling elements of a workflow process that may be nodes or edges, where nodes can be one of the following types:

- **Start and end nodes** denote the starting and ending points of the workflow.
- **A task node** represents the work to be done to achieve some given objectives.
- **A decision object** (also called condition) is used to represent alternative paths in the workflow specification depending on a conditional value(s).
- **Both task and conditions form a set of workflow graph nodes.** Each node is identified by an unique identifier.
- **A flow** defines the connection between any two objects (other than flows) in the workflow. Each flow has one source and one sink node.

In the traditional workflow graph, the whole process is defined in advance. However, our workflow is in a mixed mode, some parts of it are static that means they can be defined in advance, other parts of it may be dynamic. In order to represent the dynamic part in our workflow, we create our own symbol of dynamic task (see Fig. 2).

A customer-driven flow is shown in Fig. 3. In the traditional process, the customer enter the portal and login with an ID. They may stay for a while and buy something or leave directly. If you want to keep customers staying longer, that is to say, you have to supply personalized or one-to-one services for attracting different customers, you have to learn the user behavior. Which kind of productions the customer have bought? Which kind of productions the customer would like to buy?

Now we need to know not only past performance on the business front, but also how the customer or prospect enters our portal, aiming to target products and manage promotions and marketing campaigns. From the already demanding requirement, transaction data will be captured for further analysis.

In order to learn the user behavior better, several data mining methods will be involved. As the data mining process is invoked, the workflow will never be totally static any more.

In this case, the portal stands for not only the singe gateway for customers and enterprise marketers, but also the intelligent management system which maintain and coordinate the system applications.

If the portal needs the knowledge to determine how to promote some specific product, he will visit the knowledge base for useful knowledge as node A5 in Fig. 3. If there is no corresponding knowledge available, the mining-flow will be invoked, in this case, the targeted marketing flow will be invoked (node A6). If the mining-flow can produce some useful knowledge, the flow will come to A8. If not, the flow will end at A7 which indicates that by far there is not enough data to support the mining-flow to continue.

Before the mining result can be used to promote customer in A15, it should be evaluated first in A8. If the knowledge is good enough for use, the flow will come to A10; if the knowledge should be added with background knowledge, the flow will come to A14; if the mining result is incorrect or unsuitable, the flow will end at A13. In fact, if the mining result is incorrect, the knowledge-flow needs to revise the mining process itself to get better result next time.
4.2 Portal-driven Flow

As the portal needs to supply the personalized service for its users, the portal itself has to learn the user behavior as well. Where the visitor has been on our website? What pages has he or she visited? What is the semantic association between the pages he or she visited? Is the visitor familiar with the Web structure? Or is he or she a new user or a random one? Is the visitor a Web robot or other users? In order to do this, the portal should understand the users and make automatic modification of a website’s contents and organization to satisfy the customers.

The main process of a portal-driven flow is shown in Fig. 4. We can see that two mining-flows may be invoked at the same time, namely, the targeted marketing mining-flow and Web-log mining-flow. The mining result should be saved and evaluated, respectively. Then, different knowledge will be merged together in A11 to make automatic modification in A12.

4.3 Marketer-driven Flow

The marketer-driven flow is shown in Fig. 5. In the traditional way, the marketer can get the sales report whenever they want. However, if they want to know further, like why some kind of production is sold best, specific mining-flow, namely, ordering rule mining-flow, will be invoked.

Figure 4. Portal-driven flow

Figure 5. Marketer-driven flow

The knowledge-flow is also the application-driven flow, different mining-flow and data-flow is dynamically organized for some specific application. The main difference between the three knowledge-flows is which mining-flow is invoked for generating different knowledge and for serving specific object.

The knowledge-flows define the business process and perform the applications. By employing active knowledge, the portal can achieve personalization for:

- making a dynamic recommendation to a customer based on the customer profile and usage behavior;
- making automatic modification of a portal’s contents and organization;
- combining Web usage data with marketing data to give information about how customers use a portal for marketers.

4.4 Summary

A workflow may be instantiated several times, and several instances of the same or different workflows may be concurrently running. The workflow is not only hierarchical, but also composed. For instance, when unregistered customers want to buy some productions, the registration workflow is invoked automatically before the shopping workflow is executed.

Existing and background knowledge are stored in the knowledge base as shown in Fig. 1, which is deployed on the knowledge-grid.

Hence, as the knowledge services are deployed on the Grid platform, the dynamic workflow means dynamically

- finding, locating, organizing, and integrating of available services on the knowledge level;
- planning and organizing the data mining process on the mining level;
- finding and locating data sources on the data level.

Once the mining process is finished successfully, mining result is stored in the knowledge base. The knowledge should be evaluated before employed directly.

The evaluating process is responsible to determine:

- whether the knowledge acquired from the mining-flow is enough to use?
- whether the knowledge should be coupled with background knowledge?
- whether the mining service itself should be altered to acquire better result?
By imparting with existing knowledge, the portal could use what it knows to decide what tools and actions are appropriate for what problems and when they should be employed. In some cases, the knowledge-flow will alter the mining service itself for improving data quality. Also, the evaluating flow will invoke some mining-flows to perform a specific task.

Although we stress the importance to develop the dynamic workflow, it does not mean that all the elements in a workflow process are dynamic. From this point of view, most of the workflows are in a mixed mode which contains static elements and dynamic elements. The dynamic element is based on the mining result and result of analyzing complex conditions.

In the knowledge level, we are trying to develop a dynamic workflow management system, which governs the whole system functions.

5. Mining-flow vs. Mining-grid

The mining-flow is a collection of processes related to planning, organizing, controlling, and managing data mining process dynamically for different mining tasks on the mining-grid.

Generally speaking, customer data can be obtained from multiple customer touchpoints. In response, multiple data sources that are obtained from multiple customer touchpoints need to be integrated into a distributed data warehouse that provides a multi-faced view of their customers, their preferences, interests, and expectations for multi-aspect analysis [15].

To deal with each of data sources, separately, several data mining agents that have been developed in our group, such as:

- MVF (market values function) for learning targeted marketing values and ranking [20];
- POM (peculiarity-oriented Web-log mining) for finding peculiarity data/rules [21];
- LOI (learning with ordered information) for discovering ordering rules/important attributes [12].

are employed to analysis three data sources on the lower level, and support the knowledge services on the upper level.

In the mining-grid, mining agents are deployed on the related datasets as mining services. For instance, the MVF agent is deployed on the customers DB; the POM agent is deployed on the Web-logs DB; and the LOI agent is deployed on the products DB.

From the top-down perspective, different data mining methods are deployed on the mining-grid as mining services. Three knowledge-flows on the knowledge level will evoke corresponding mining services with standard interfaces when needed (see Figs. 3, 4 and 5). However, in the mining level, different mining methods work just like agents, that is to say, they are working in a self-organizing way.

5.1 Multiple Data Mining Agents

One of the main reasons for developing multiple data mining agents is that we cannot expect to develop a single data mining algorithm that can be used to solve all problems since complexity of the real world applications. Hence, various data mining agents need to be cooperatively used in the multi-step data mining process for performing multi-aspect analysis as well as multi-level conceptual abstraction and learning.

The other reason for developing multiple data mining agents is that when performing multi-aspect analysis for complex problems, a data mining task needs to be decomposed into sub-tasks. Thus these sub-tasks can be solved by using one or more data mining agents that are distributed over different computers and multi-data repositories on the Grid. Thus the decomposition problem leads us to the problem of distributed cooperative system design.

5.2 Data Mining Process

In the mining level, data mining process is also the mining-flow. The mining-flows may be transferred, instantiated, and executed several times in the distributed environment, and several instances of the same or different mining-flows may be running concurrently.

Hence, one of the key questions in the mining level is how to dynamically plan the mining process on the mining-grid. That is to say, how to find, locate, integrate, and use the distributed resources for mining process.

Figure 6 is an example of mining-flow, i.e., MVF mining-flow. MVF agent involves the identification of customers having potential market value by studying the customers’ characteristics and the needs, and selects certain customers to promote.

Hence, when the MVF process is employed, it will employ data-flow first to get the appropriate data sources to analysis, and when the mining process is finished, the mining result, i.e., market value, will be inserted into the corresponding dataset. Whether this mining result is good enough for use or not will left for evaluation in the knowledge level.

These kinds of mining applications can be built by reusing capabilities provided by existing grid-enabled Web services. The difference between the mining-flows, i.e., POM and LOI mining-flows, is that different mining methods and data-flows are involved.
5.3 Data Quality Analysis

The other key question in the mining level is to determine when and how to start the mining processes. We use data quality to solve this kind of problem. Hence, the mining-flow will start in two situations, one is invoked by knowledge-flow from knowledge level; the other is the result of data quality analysis is satisfied.

The mining-flow can be invoked by the knowledge-flow from knowledge level (see Figs. 3, 4 and 5). However, most of mining-flows are generated by the result of analyzing data quality. A data quality workflow is shown in Fig. 7.

Data quality analyzing services are deployed on the mining-grid, which can keep watching the data automatically. If the mining status is not satisfied, the mining-flow will not be invoked, see A3 in Fig. 7. If the mining status is satisfied, the mining-flow will be invoked, see A4 in Fig. 7.

For instance, the MVF mining-flow will not be started until we have enough members in positive and negative datasets. If we want to know who is interested in our portal. A data quality analyzing service will keep watching the customer related datasets, until the following conditions is satisfied.

- We have enough members in the customer dataset \( U \),
- We have enough members of registration customers \( P \),
- We have enough members in the negative dataset \( D \).

In the above situations, the mining process is almost the same, the difference are the data quality items. Based on the different items to analysis, corresponding data quality services should be developed.

For instance, when we want to know who is interested in certain category (sports, music, or book). Based on the different category, positive and negative member needed for starting TM is totally different. In this case, we can’t give this value in advance. We can only set a primitive number from experience, the knowledge level will analyze and evaluate the result after each promotion based on the real effect.

And then revise the data quality value. Hence, data quality analysis is a self-learning process.

When we try to analyze from another angle like suppliers, for instance, ranking the same kind of production supplied by different companies or different kinds of productions supplied by the same company, another ordering rule [14] is needed. This ordering rule is formulated as the process of finding associations between orderings on attribute values and the overall ordering of objects.

For different mining methods, like ordering rule and Web-log mining, data quality analysis will be totally different. Hence, developing a appropriate and effective method for data quality analysis is a main challenge for mining level.

6. Data-flow vs. Data-grid

The data-flow is a collection of descriptions for the dynamic relationship among multiple data sources on the data-grid.

From the beginning, we only have the products DB. The customers and Web-log information are accumulated during the business activities such as registration, searching/browsing and shopping. Customers visit our portal, make registration and browse. Finally, they may buy something they are interested in or leave directly.

When a customer buy some production, his information will be recorded in both customers and products DB. Also the relationship between the data sources will be created and recorded. Also, different mining-flows will dynamically build different relationship between the data sources.

Typically, there are four kinds of relations between the three data sources, namely solid relation, strong relation, weak relation, and uncertain relation.

If customers have bought some kinds of productions, a solid relation is built between the customers DB and products DB. An MVF mining-flow may build strong relation between the customers DB and products DB, which indicates some customer has strong possibility to buy some production. A Web-log mining-flow always build the weak relation between the data sources. Because Web-log mining usually
give the orientation of a group of users. Uncertain relation indicates so far there is not enough data to build the relation between the data sources.

Data storing and retrieving are deployed on the Grid platform, like Globus, as a standard Grid service. OGSA-DAI is used to build database access applications [23].

The aim of OGSA-DAI is to provide the middleware glue to interface existing databases, other data resources and tools to each other in a common way based on the Open Grid Services Architecture (OGSA). This middleware is based on the GGF-defined OGSI specification and layered on top of the Globus Toolkit 3 OGSI implementation (GT3 Core).

Multiple data sources are recorded to the corresponding database through Grid service on the distributed sites. For instance, customer related databases and product related databases are deployed on the different sites. Every touch points of customers will be reordered in the Web-logs DB, which is deployed near the portal for the time-saving purpose.

A typical OGSA-DAI process is shown in Fig. 8. The portal first has to discover the data resource by sending a call the registry (1a), the DAISGR (Database Access and Integration Service Group Registry) registry responds with a handle (1b) which tells it about the Factory. The portal then asks the GDSF (Grid Data Service Factory) to create a Grid Data Service for it to interact with the data source (2a), and the Factory does so (2b) and returns a handle to the new service (2c). The portal then sends its query to the GDS (Grid Data Service) in step 3a, which executes a query and returns the result directly to the portal. Once the GDS has been created, the portal can send multiple queries to it without any further interaction with the registry or Factory.

Multiple data-flows on the data grid is shown as Fig. 9.\footnote{Derived from [2]}

vices. G indicates deployed OGSA-DAI services which will supply the data service registry to locate the different database.

For instance, item 1 in Fig. 9, a workflow from upper level query the registry first, then get the status and result of the database. Item 2 indicates that workflow can tell the OGSA-DAI registry to transfer the database result to other workflows. Item 3 indicates that workflow can transfer the database registry information to other workflow directly.

Also, if the workflow get the database registry information, he can finish updating and inserting operations as well as querying (see items 4, 5 and 6 in Fig. 9).

Items 7, 8 and 9 in Fig. 9 show the relationship between the three data sources. A workflow can get the different registry information about the database, i.e., G1 to customer database, G2 to production database. R1 indicates the solid or strong relationship between them. Different relationship is built based on which workflow is invoked.

Different data on the different sites will help to deal with the scalability and complexity of real world, efficiently and effectively. Multiple databases are deployed on the distributed sites, which will help to lower the workload for the server, for instance, the portal. However, this kind of distribution will not let the customers feel the delay obviously. As the customer will not like to visit a slow response website.

Multiple data sources obtained from multiple customer touch points need to be preprocessed and integrated for mining services.

7. Conclusions

The underlying Grid resource set is typically heterogeneous even within the same class and type of resource. Fundamental to both resource management and integration
is the ability to discover, allocate, negotiate, monitor, and manage the use of network-accessible capabilities in order to achieve various end-to-end or global qualities of service [6].

The proposed workflow management system offers higher level toolkits and techniques for dynamically organizing distributed mining process and extracting knowledge from datasets available on the Grid. The development of such a system is the main goal of our research.

This paper illustrates how to use the multi-level conceptual model for building an e-Business portal, the most common challenge is how to deploy and use multiple data sources on the Grid and how to present the complex relationship between them. How to build effective data quality analysis methods for multiple mining processes is still needed to be discussed.

The development of e-Business portals is based on the paradigm of Wisdom Web computing in which AI and advanced IT are incorporated to make a reality of intelligent portals. The ultimate goal is to develop plug and play application package that serves the users wisely for e-Business activities.

References


