A WORKFLOW BASED APPROACH FOR KNOWLEDGE GRID APPLICATION

Marcello Castellano and Cesaria Digregorio
Dipartimento di Elettrotecnica ed Elettronica, Politecnico di Bari, via E. Orabona, 4 - 70125, Bari, Italy

Keywords: Grid Computing, Middleware, Workflow, Knowledge Discovery, Knowledge Grid Workflow, Knowledge Grid Workflow Management System.

Abstract: An expanding area of study regards the development of computing platforms able to provide easy online access to geographically distributed data collections with the aim of producing knowledge synthesis from that data. Computational grid technology can form a computing infrastructure which can assist in providing a platform for distributed computing. In this paper is presented a case study to propose a Knowledge-Grid Workflow Management System as a middleware solution to develop a Knowledge Grid application.

1 INTRODUCTION

The first decade of the new millennium has been called the *Data Decade* due to the fact that data growth is surpassing computational growth, and many of the more important advances in science and engineering originate from analyses and syntheses of large data collections. Biomedical images, protein deposits, digital maps and research into the building blocks of matter are just a few examples of data collections that could furnish raw data to online searches and data analyses (F. Berman, 2001). The activity that must be carried out on the data is connected to the way in which they are retrieved and to the ability to analyse them in order to create a synthesis or, better yet, to transform the data into something useful, knowledge for the user. Hence, the development of computing platforms able to furnish easy online access to geographically distributed data collections with the aim of producing scientific knowledge through simulation, phenomena modelling, and the results of data analyses is an area undergoing rapid expansion.

There are several methods, strategies and techniques for accessing and synthesising knowledge from data as mining and inference. However, satisfying requests for knowledge often involves coordinating several tasks and orchestrating different contributions, communications and components of information retrieval and syntheses (S. Sarnikar, 2007). Today, grid technologies are the most promising answer for the creation of applied scenarios with scalable performance and a large capacity for sharing both physical and logical computing resources. A Knowledge-Grid is the convergence of a comprehensive computational infrastructure with scientific data collections and applications for routinely supporting the synthesis of knowledge from that data. Grid Computing applications can become truly operative through the development of a middleware which operates between the grid system and the user's application.

In this paper, we propose the use of a Knowledge-Grid Workflow Management System as a middleware which both allows the user to specify the model of knowledge synthesis desired and transparently implement it, distributing the workload suitably onto the Knowledge-Grid system in use. In particular section 2 briefly discusses concepts as knowledge and related synthesis techniques, grid systems, the Knowledge-Grids and workflow. Section 3 presents functional models of Grid Workflow and Knowledge-Grid Workflow. Section 4 discusses a case study which presents a workflow model for knowledge synthesis applied to text mining to automatically extract data related to symptoms and pathologies from a repository of scientific documents.

Castellano M. and Digregorio C. (2009).
A WORKFLOW BASED APPROACH FOR KNOWLEDGE GRID APPLICATION.
In *Proceedings of the International Conference on Knowledge Management and Information Sharing*, pages 230-235
DOI: 10.5220/0002331702300235

2 BACKGROUND

In general, knowledge can be experience, concepts, values, or beliefs that increase an individual's ability to take effective action (F. Zheng, 2008). Knowledge can be either implicit or explicit. The former is represented by tacit experience which can come through individual ideas, intuition, experience, values and judgements. This type of knowledge is dynamic in nature. It can be accessed only through direct participation and communication with field experts that possess this knowledge. The know-how of each knowledge worker is accordingly based on this tacit (or implicit) knowledge. Instead, explicit knowledge usually includes anything that is saveable in an electronic format or in other words what we are able to transcribe and to share.

Knowledge discovery can be defined as "the non-trivial extraction of implicit, unknown, and potentially useful information from data. When working with texts, knowledge discovery refers generally to the process of extracting interesting information from a large amount of unstructured textual documents. The goal of this process is to find and extract useful patterns. To do this, specific methods and algorithms from the fields of machine learning and statistics are applied. Text mining is, thus, the application of these algorithms and methods to texts "(U.M. Fayyad, 1996).

Grid is an infrastructure which allows shared resources to be coordinated inside dynamic organisations, be they individuals, institutions or resources. It offers a flexible environment where resources can be dynamically reorganised without altering any active processing on the GRID and provide connectivity for data distributed in different locations. This can resolve transparency problems related to location while providing a mechanism which allows easier access to and management of distributed data as well as the virtualisation and sharing of GRID connected resources. To manipulate intensive computation procedures, the platform can provide automatic allocation of resources, scheduling and algorithm implementation in relation to the availability, capacity and position of these resources. A GRID can increase efficiency while reducing the cost of computational networks by decreasing data processing times, optimizing resources, and distributing the workload. Thus, users are provided the results of large operations with greater speed and lower costs (I. Foster, 2001).

Attempts to automate knowledge processes date to the early 1980s. Several processes have been employed on parallel computing platforms to achieve high performance on the analysis of large data sets stored on a single site. Recently, the demand for knowledge processes has expanded to include the management and analyses of multi-site and multi-owner data repositories. This task involves large data-sets, the geographic distribution of data, users and resources, and computational intensive analysis demands for new parallel and distributed platforms for knowledge processes as computational grid technology. The resulting application of grid technology to the knowledge field has been termed Knowledge-Grid (M.Cannataro, 2001).

Workflow automation technology has been developed to facilitate organizational coordination and collaboration by automating entire work processes and controlling the flow of information among participants. A workflow can be used to define the work process, control activity requests, route relevant documents to the appropriate agents, enforce deadlines, and monitor the progress of work (S. X. Sun, 2008). The Workflow Management Coalition (WFMC) defines a workflow as "… the total or partial automation of business procedures where documents, information or tasks are passed between participants according to a defined set of rules …" (www.wfms.com). A business process is a group of necessary tasks and a set of conditions which determines the order of their completion. A task is a logical unit of work that must be performed by a resource in its entirety. A resource can be a person or machine or it can be a group of persons or machines that perform specific tasks. The performance of a task by a resource is called an activity (Wil van der Aalst, 2002).

Hence, a workflow can be seen as a structure which not only contains tasks/activities but also coordinates and supervises their execution.

Different types of workflow have been identified:

Collaborative workflows manage less rigid processes and allow connections among those users closest to the collaboration as well as work groups;

Structured workflows manage structurally well-defined and repeatable activities which can be specified through a series of rules. Examples of structured workflows are: (a) administrative workflows which manage the flow of electronic forms, integrating them with message systems or email; and (b) production workflows which manage the flow of well-structured work, defined by well-formalised rules and dependencies;

Ad Hoc workflows are created by using lighter systems which give the user the task of identifying the correct procedural steps to take each time a

dynamic modification is required in the performance procedure.
In other words, workflows are a specific kinds of processes, whose transitions between activities are controlled by an information system called a Workflow Management System (E. A. Stohr, 2001).

Workflow Management Systems (WfMS) are a new type of computer technology which aid the automation of business procedures. More specifically, a WfMS defines, manages and performs workflow through the use of software (www.wfms.com). This software can also be considered a middleware which connects an organisation's offices and legacy applications (E. A. Stohr, 2001). A WfMS forms a bridge between those who are required to develop the given tasks and the computer applications which are the operating tools used by the people to develop each activity. It has been designed to produce work which can be monitored and to encourage communication among the different parts [Wil van der Aalst, 2002].
WfMSs can be schematically outlined in three functional areas (Fig. 1):

1. The built-time modelling operation which defines and eventually modifies the task or activity assigned.
2. The runtime workflow enactment service operation which manages workflow procedures in an operative environment and orders the various activities to be managed as part of a procedure.
3. The interactions amongst the runtime, the human user and the IT application tools in order to control the activities.

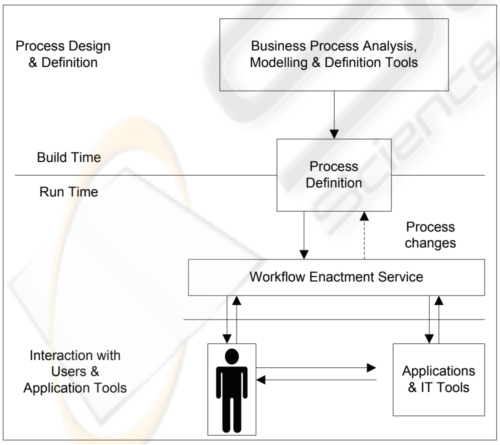

Figure 1: Workflow Management System characteristics.

The ability to distribute tasks and information between participants is a major distinguishing feature of workflow runtime infrastructure. The distribution function may operate at a variety of levels (workgroup to inter-organisation) depending upon the scope of the workflows. It may also use a variety of underlying communications mechanisms (electronic mail, message passing, distributed object technology, etc) (Fig. 2) (www.wfms.com).

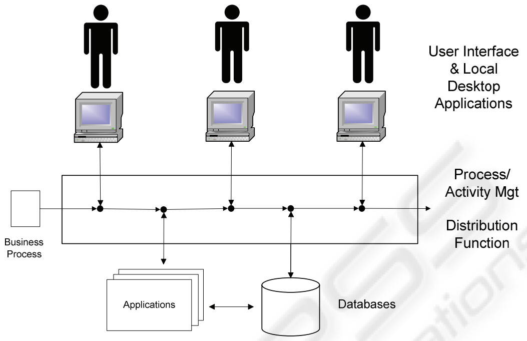

Figure 2: Distribution within the workflow enactment service.

WFMSs bring with them a series of advantage like increased efficiency, reduced time and errors, improved control over procedures through the standardisation of work methods, and the availability of verification tools.

3 KNOWLEDGE-GRID WORKFLOW

In recent years, several techniques for workflow management have been developed, especially, in the business field and any of these approaches have been investigate for Grid Workflow applied in scientific fields (J. Yu, 2005). A Grid Workflow is a collection of tasks processed in a well-defined order on distributed resources. WfMSs which define, manage and perform workflow for Grid applications are constantly expanding. The systems offer many advantages: the possibility of dynamically building applications which take advantage of distributed resources; the use of resources located in different areas to improve data treatment or reduce the performance costs; the ability to perform extended runs in different administrative areas to obtain specific procedure capacities; the integration of groups involved in the management of different aspects of an experiment which make up a workflow, encouraging collaboration between organisations (J. Yu, 2005).
Figure 3 shows the functional architecture and features of components of the Grid-based workflow system model proposed by the WfMC. The user

interacts with tools that produce a workflow model (Workflow Design & Definition) which then undergoes the implementation service (Workflow Execution & Control). This service provides a series of functions as scheduling, fault tolerance, and data movement. It is built on a Grid middleware through which the Grid resources can be used. Both in the model building phase and in that of execution, the resource and application information is retrieved by examining the Grid services dedicated to the information (J.Yu, 2005).

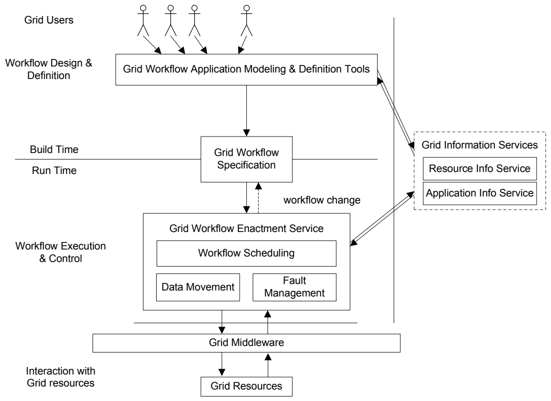
Grid Users
Workflow Design & Definition
Grid Workflow Application Modeling & Definition Tools
Build Time
Run Time
Grid Workflow Specification
workflow change
Grid Information Services
Resource Info Service
Application Info Service
Workflow Execution & Control
Grid Workflow Enactment Service
Workflow Scheduling
Data Movement
Fault Management
Grid Middleware
Interaction with Grid resources
Grid Resources

Figure 3: Grid Workflow Management System.

Currently, knowledge sharing and knowledge transfer processes require the manual integration and organization of information channels supported by a fragmented set of tools such as email, discussion forums, search engines, etc. The lack of an approach that can model knowledge flows and coordinate the working of various tools to automate them is a major drawback and deterrent to knowledge sharing and improved knowledge worker productivity in organizations (Bruno 2002; Moore and Rugullies 2005).

The automation of knowledge flow requires the integration of information retrieval mechanisms with workflow systems. While information retrieval mechanisms provide discovery and matching services, workflow systems coordinate the call for the appropriate intelligent service and automate the routing and delivery of messages and documents. Specifically, given a set of user specified constraints, a workflow can be designed so that it can automatically invoke intelligent services, contact experts, retrieve documents and present the results to the user (S. Sarnikar, 2007).

A knowledge workflow has several unique features when compared to a generic workflow. For example, the number of parallel activities needed to locate experts and databases is decided dynamically at runtime and is not modelled a priori as in typical production workflows. Similarly, the user's choice of activities following a response cannot be modelled statically beforehand and is instead represented by a 'model update point' in the knowledge workflow.

Although knowledge workflows are similar to structured business procedures in some aspects, there are some main differences to perform them using existing WfMSs. The first problem is that in a typical business procedures the processing of orders, the flow of control and the data flow are all predetermined. Knowledge flow is "ad hoc" by nature and evolves on the basis of user requests and on the limits of systems that are difficult to define a priori. The second difficulty is that the assemblage of Knowledge flow models is carried out at the moment of execution. Thus, it is necessary to combine the process modelling and the workflow management enactment function with the information modelling and delivery mechanisms of knowledge management (E. A. Stohr, 2001).

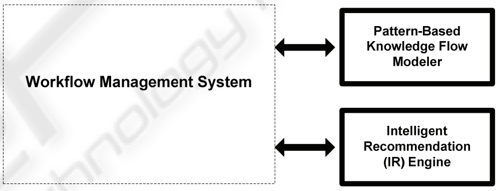
Workflow Management System
Pattern-Based Knowledge Flow Modeler
Intelligent Recommendation (IR) Engine

Figure 4: KWMS architecture.

A Knowledge Workflow Management System (KWMS) can be implemented by extending the WfMC architecture of workflow management systems with two knowledge workflow specific components: (1) a pattern-based knowledge flow model generator and (2) an IR System (see fig. 4). The knowledge workflow modeller is used to develop knowledge workflow models by assembling basic knowledge workflow patterns to meet the requirements of a specific knowledge request. The IR

System engine consists of an information retrieval engine and an expertise locator engine. The information retrieval engine is used to execute functions such as document recommendation, aggregation, filtering and other retrieval related functions (S. Sarnikar, 2007).

4 A CASE STUDY

The Knowledge-Grid Workflow application here taken into account is designed to extract new and useful information about symptoms and pathologies from a large collection of unstructured scientific documents (M.Castellano, 2009). This is a frequent study activity in biomedicine called bio-entity recognition and it is receiving ever greater attention. Bio-entity recognition aims to identify and classify technical terms corresponding to the instances of concepts that are of interest to molecular biologists. Examples of such entities include the names of proteins, genes, their location of activity (i.e., the names of cells or organisms), drugs, symptoms, pathologies and so on. Entity recognition is becoming increasingly important with the massive increase in reported results due to high throughput experimental methods (H. Shatkay, 2003).

1. The first workflow step is the creation of a Textual Data Repository from unstructured texts such as scientific medical publications, abstracts of clinical articles, or parts of texts concerning health services. Mining large document collections requires pre-processing the text documents and storing the information in a data structure since the latter is more appropriate for further processing than a plain text file. Once a Textual Data Repository is created, the Knowledge Discovery in Text technique can be applied. The Knowledge Discovery in Text process is consists of *Text Refining* and *Text Mining* phases.
2. Text Refining input are not-structured data such as texts or semi-structured data like HTML pages. This next step requires dividing the documents into a stream of words by removing all punctuation marks and by replacing tabs and other non-text characters with single white spaces. Articles, conjunctions, prepositions, etc. must be removed from the documents. *Text Refining* output can be stored in a database, XML file or in other structured forms which are referred to as an Intermediate Form and Text Mining techniques are applied to this Form.

At this point, our Knowledge Workflow is broken down into three Text Mining sub phases: *document clustering*, *document categorization*, and *pattern extraction.*

3. Document clustering is the assignment of a multivariate entity to a few categories (classes, groups) previously undefined. The goal is to gather together similar entities. Textual clustering is used as an automatic process which divides a collection of documents into groups. Inside these groups, the documents have similarities based on selected characteristics: author, length, dates, and keywords. Textual clustering can be used to provide a planning of the contents of document collections, to identify hidden similarities, to facilitate the process of browsing and to find correlated or similar information. If the clustering works with keywords or features that represent the semantics of the documents, the identified groups will be distinguished on the basis of the different topics being discussed in the corpus.
4. Document categorisation requires that the objects must be attributed to one or more class or category which will already have been identified. Classification is the process in which meaningful correlations among frequent data are identified. There are association rules for Text Categorization. All algorithms operate in two phases to produce association rules. First, all the whole keywords with greater or equal support with respect to the reference are listed to create what is called the frequent set. Then, all the association rules, that can be derived from the frequent set and that satisfy the given confidence, are established.
5. Pattern Extraction is the identification of some patterns following the analysis of associations and tendencies. The discovery of associations is the process in which meaningful correlations among frequent whole data are found.

At the end of Text Mining, the analysis of a large collection of biomedical papers will allow us to extract biological entities related to symptoms and pathologies. The Knowledge-Grid Workflow described above is illustrated in fig.5:

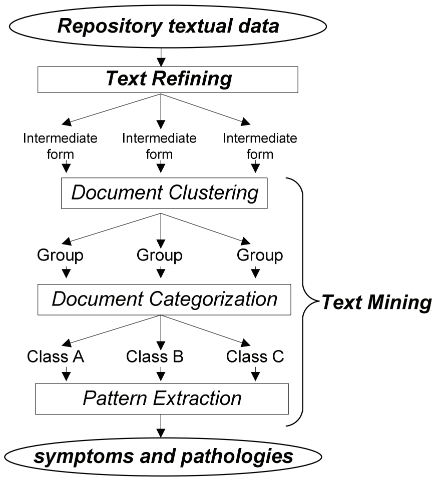

Figure 5: Example of a Knowledge-Grid Workflow used for Symptom and Pathology Discovery.

Application problems often occur in data-intensive situations based on knowledge methods. These circumstances require that the same logic be applied to a large collection of different data which are independent from each other. Hence, the limitations will be technological if these problems are addressed with traditional machines that sequentially perform the same set of instructions on an entire collection of homogeneous and independent data. The time required for execution will also increase according to the size of the collection and this will become the limiting factor in these applications (M.Castellano, 2009). Therefore, the creation and management of a dynamic, multi-step workflow is necessary. Our workflow model represents a knowledge synthesis process concerning the discovery of terms which describe medical symptoms and pathologies. It can be generated and executed by a Knowledge-Grid Workflow Management System which automated both the model building procedure as well as its execution, guided by the Grid technology. In particular, by retrieving the Grid Workflow architecture, it was possible to interact with tools that first generated a workflow model (Workflow Design & Definition) and then underwent an execution service (Workflow Execution & Control). When the Knowledge Workflow model is being defined and also during the execution phase, the Grid services can be examined and the availability of resources can be verified. It should be noted that the resources allowed us to perform the various phases of the Knowledge Workflow or, better, the phases of Text Refining and Text Mining which resulted in the extraction of the terms describing symptoms and pathologies.

5 CONCLUSIONS

The development of computing platforms able to provide easy online access to geographically distributed data collections with the aim of producing knowledge synthesis from that data is showing an increasing interest. The complexity of the knowledge process find the operational answer through the use of workflow technologies. In this paper has been presented a case study based on automatic extraction of symptoms and pathologies in order to expose the workflow approach for knowledge-Grid. The Knowledge-Grid Workflow Management System is a suitable middleware that allow both an high level design of the whole Knowledge process and the execution of it having a set of tools already configured and running in a Grid based environment.

REFERENCES

Fran Berman, 2001. From TeraGrid to Knowledge Grid. Comunication of the ACM, Vol. 44, No 11.

Surendra Sarnikar, J. Leon Zhao, 2008. Pattern-based knowledge workflow automation: concepts and issues. *Information Systems and E-Business Management,* 6(4), pp 385-402.

Fanghua Zheng，Guojie Zhao, 2008. New Perspective of Knowledge Management based on Dual dimension

Classification of Knowledge. *International Symposium on Knowledge Acquisition and Modeling.*

Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. 1996. From Data Mining To Knowledge Discovery: An Overview. In Advances In Knowledge Discovery And Data Mining, eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA., pp. 1-34.

I. Foster, C. Kesselman, Steven Tuecke, 2001. The Anatomy of the Grid. Enabling Scalable Virtual Organizations. *International Journal of Supercomputer Applications.*

M. Cannataro, D.Talia, P. Trunfio, 2001. KNOWLEDGE GRID: High Performance of Knowledge Discovery on the Grid. *Proceedings of the Second International Workshop on Grid Computing.*

Sherry X. Sun, J. Leon Zhao, 2008. Developing a Workflow Design Framework Based on Dataflow Analysis. *Proceedings of the 41st Hawaii International Conference on System Sciences*.

www.wfms.com

Wil van der Aalst, Kees van Hee, 2002. Workflow Management: Model, Methods, and System. The MIT Press Cambridge, Massachusetts London, England.

Edward A. Stohr, J. Leon Zhao, 2001. Workflow Automation: Overview and Research Issues. *Information Systems Frontiers 3:3, 281–296.*

Jia Yu, Rajkumar Buyya, 2005. A taxonomy of Workflow Management System for Grid Computing. *Journal of Grid Computing, Volume 3, Numbers 3-4, Pages: 171-200*, Springer Science+Business Media B.V.

M.Castellano, G.Mastronardi, R.Bellotti, G.Tarricone, 2009. A bioinformatics Knowledge Discovery Application for Grid computing. *EMBnet Conference.*

Shatkay H, Feldman R., 2003. Mining the biomedical literature in the genomic era: an overview. *Journal of Computational Biology*, 10 (Suppl 6):821-855.