Encyclopedia of Data Warehousing and Mining

Second Edition

John Wang
Montclair State University, USA

Volume III
K–Pri

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content:	Kristin Klinger
Director of Production:	Jennifer Neidig
Managing Editor:	Jamie Snavely
Assistant Managing Editor:	Carole Coulson
Typesetter:	Amanda Appicello, Jeff Ash, Mike Brehem, Carole Coulson, Elizabeth Duke, Jen Henderson, Chris Hrobak, Jennifer Neidig, Jamie Snavely, Sean Woznicki
Cover Design:	Lisa Tosheff
Printed at:	Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of data warehousing and mining / John Wang, editor. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
Summary: "This set offers thorough examination of the issues of importance in the rapidly changing field of data warehousing and mining"--Provided by publisher.
ISBN 978-1-60566-010-3 (hardcover) -- ISBN 978-1-60566-011-0 (ebook)
1. Data mining. 2. Data warehousing. I. Wang, John,
QA76.9.D37E52 2008
005.74--dc22
2008030801

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this encyclopedia set is new, previously-unpublished material. The views expressed in this encyclopedia set are those of the authors, but not necessarily of the publisher.

If a library purchased a print copy of this publication, please go to http://www.igi-global.com/agreement for information on activating the library's complimentary electronic access to this publication.

Path Mining and Process Mining for Workflow Management Systems

Jorge Cardoso
SAP AG, Germany

W.M.P. van der Aalst
Eindhoven University of Technology, The Netherlands

INTRODUCTION

Business process management systems (Smith and Fingar 2003) provide a fundamental infrastructure to define and manage business processes and workflows. These systems are often called process aware information systems (Dumas, Aalst et al. 2005) since they coordinate the automation of interconnected tasks. Well-known systems include Tibco, WebSphere MQ Workflow, FileNet, COSA, etc. Other types of systems, such as ERP, CRM, SCM, and B2B, are also driven by explicit process models and are configured on the basis of a workflow model specifying the order in which tasks need to be executed.

When process models or workflows are executed, the underlying management system generates data describing the activities being carried out which is stored in a log file. This log of data can be used to discover and extract knowledge about the execution and structure of processes. The goal of process mining is to extract information about processes from logs.

When observing recent developments with respect to *process aware information systems* (Dumas, Aalst et al. 2005) three trends can be identified. First of all, workflow technology is being embedded in service oriented architectures. Second, there is a trend towards providing more flexibility. It is obvious that in the end business processes interface with people. Traditional workflow solutions expect the people to adapt to the system. However, it is clear that in many situations this is not acceptable. Therefore, systems are becoming more flexible and adaptable. The third trend is the omnipresence of event logs in today's systems. Current systems ranging from cross-organizational systems to embedded systems provide detailed event logs. In a service oriented architecture events can be monitored in various ways. Moreover, physical devices start to record events. Already today many professional systems (X-ray machines, wafer stepper, high-end copiers, etc.) are connected to the internet. For example, Philips Medical Systems is able to monitor all events taking place in their X-ray machines.

The three trends mentioned above are important enablers for path mining and process mining. The abundance of recorded events in structured format is an important enabler for the analysis of run-time behavior. Moreover, the desire to be flexible and adaptable also triggers the need for monitoring. If processes are not enforced by some system, it is relevant to find out what is actually happening, e.g., how frequently do people deviate from the default procedure.

BACKGROUND

Path mining can be seen as a tool in the context of Business Process Intelligence (BPI). This approach to path mining uses generic mining tools to extract implicit rules that govern the path of tasks followed during the execution of a process. Generally, the realization of a process can be carried out by executing a subset of tasks. Path mining is fundamentally about identifying the subset of tasks that will be potentially be triggered during the realization of a process. Path mining is important to process Quality of Service (QoS) prediction algorithms (Cardoso, Miller et al. 2004). In processes for e-commerce, suppliers and customers define a contract between the two parties, specifying QoS items such as products or services to be delivered, deadlines, quality of products, and cost of services. A process, which typically has a graph-like representation, includes a number of linearly independent control paths (i.e. paths that are executed in parallel). Depending on the path followed during the execution of a process, the QoS may substantially be different. If we can predict with a certain degree of confidence the path that will be followed at runtime, we can significantly increase the precision of QoS estimation algorithms for processes.

Process mining has emerged as a way to analyze systems and their actual use based on the event logs they produce (Aalst, Dongen, et al. 2003; Aalst, Weijters, Maruster, 2004; Aalst, Reijers, et al. 2007). Note that, unlike classical data mining, the focus of process mining is on concurrent processes and not on static or mainly sequential structures. Process mining techniques attempt to extract non-trivial and useful information from event logs. One element of process mining is control-flow discovery, i.e., automatically constructing a process model (e.g., a Petri net) describing the causal dependencies between activities. In many domains, processes are evolving and people, typically, have an oversimplified and incorrect view on the actual business processes. Therefore, it is interesting to compare reality (as recorded in the log) with models. Since process mining is so important for organizations, there was the need to develop a system which implements the most significant algorithms developed up to date. Therefore, the ProM framework has been developed as a completely plug-able environment. Different research groups spread out over the world have contributed to ProM. Currently there are more than 150 plug-ins available, thus supporting all aspects of process mining (Aalst, Reijers, et al. 2007).

SETTINGS

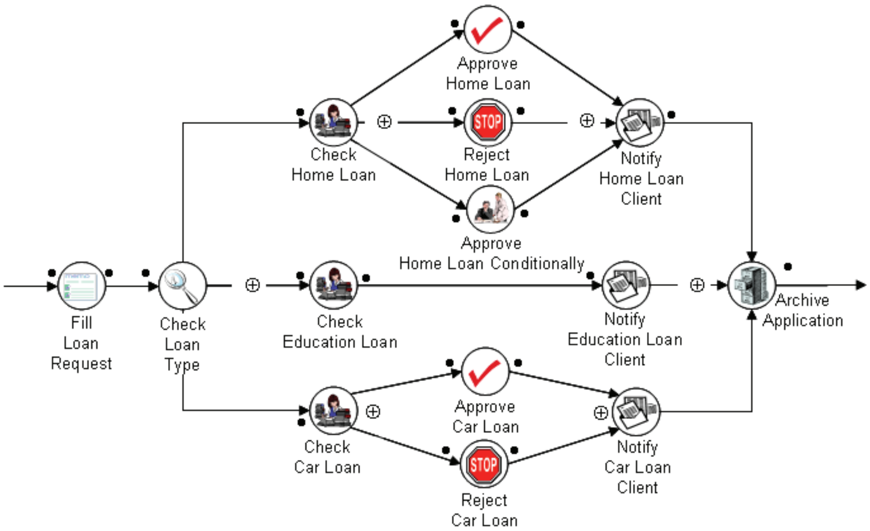
This section presents a typical business process model and illustrates also a typical process log. These two elements will be use to explain the concepts of path mining and process mining in the next section.

Business Process Scenario

A major bank has realized that to be competitive and efficient it must adopt a new and modern information system infrastructure. Therefore, a first step was taken in that direction with the adoption of a workflow management system to support its business processes. All the services available to customers are stored and executed under the supervision of the workflow system. One of the services supplied by the bank is the loan process depicted in Figure 1.

The process of the scenario is composed of fourteen tasks. For example, The Fill Loan Request task allows clients to request a loan from the bank. In this step, the client is asked to fill in an electronic form with personal information and data describing the condition of the loan being requested. The second task, Check Loan Type, determines the type of loan a client has requested

Figure 1. The loan process



and, based on the type, forwards the request to one of three tasks: Check Home Loan, Check Educational Loan, or Check Car Loan.

Process Log

Many systems have some kind of event log often referred to as “history”, ”audit trail”, ”transaction log”, etc. (Agrawal, Gunopulos et al. 1998; Grigori, Casati et al. 2001; Sayal, Casati et al. 2002; Aalst, Dongen et al. 2003). The event log typically contains information about events referring to a task and a case. The case (also named process instance) is the “thing” which is being handled, e.g., a customer order, a job application, an insurance claim, or a building permit. Table 1 illustrates an example of a process log.

PATH MINING

To perform path mining, current process logs need to be extended to store information indicating the values and the type of the input parameters passed to tasks and the output parameters received from tasks. Table 2 shows an extended process log which accommodates input/output values of tasks parameters that have been generated at runtime. Each ‘Parameter/Value’ entry has a type, a parameter name, and a value (for example, string loan-type=”car-loan”).

Additionally, the process log needs to include path information, a path describing the tasks that have been executed during the enactment of a process. For example, in the process log illustrated in Table 2, the service NotifyUser is the last service of a process. The log has been extended in such a way that the NotifyUser record contains information about the path that has been followed during the process execution.

Process Profile

When beginning work on path mining, it is necessary to elaborate a profile for each process. A profile provides the input to machine learning and it is characterized by its values on a fixed, predefined set of attributes. The attributes correspond to the task input/output parameters that have been stored previously in the process log. Path mining will be performed on these attributes.

Profile Classification

The attributes present in a profile trigger the execution of a specific set of tasks. Therefore, for each profile previously constructed, we associate an additional attribute, the path attribute, indicating the path followed

Table 1. Process log

Date	Process	Case	Task	Task instance	Cost	Dur.	…
6:45 03-03-04	LoanApplication	LA04	RejectCarLoan	RCL03	$1.2	13 min	…
6:51 03-03-04	TravelRequest	TR08	FillRequestTravel	FRT03	$1.1	14 min	…
6:59 03-03-04	TravelRequest	TR09	NotifyUser	NU07	$1.4	24 hrs	…
7:01 03-03-04	InsuranceClaim	IC02	SubmitClaim	SC06	$1.2	05 min	…
…	…	…	…	…	…	…	…

Table 2. Extended process log

…	Case	Task	Task instance	Parameter/Value	Path	…
…	LA04	NotifyCLoanClient	NLC07	string e-mail=”jf@uma.pt”	…	…
…	LA05	CheckLoanRequest	CLR05	double income=12000; string Name=”Eibe Frank”;	…	…
…	TR09	NotifyUser	NU07	String e-mail=jf@uma.pt; String telef=”35129170023”	FillForm->CheckForm-> Approve->Sign->Report	…
…	…	…	…		…	…

when the attributes of the profile have been assigned to specific values. The path attribute is a target class. Classification algorithms classify samples or instances into target classes. Once the profiles and a path attribute value for each profile have been determined, we can use data mining methods to establish a relationship between the profiles and the paths followed at runtime.

Experiments

In this section, we present the results of applying an algorithm to a synthetic loan dataset. To generate a synthetic dataset, we start with the process presented in the introductory scenario, and using this as a process model graph, log a set of process instance executions.

The data are lists of event records stored in a process log consisting of process names, instance identification, task names, variable names, etc.Table 3 shows the additional data that have been stored in the process log. The information includes the task variable values that are logged by the system and the path that has been followed during the execution of instances. Each entry corresponds to an instance execution.

Process profiles are characterized by a set of six attributes: income, loan_type, loan_amount, loan_years, name and SSN. These attributes correspond to the task input/output parameters that have been stored previously in the process log presented inTable 3. Each profile is associated with a class indicating the path that has been followed during the execution of a process when the attributes of the profile have been assigned specific values. The last column ofTable 3 shows the class named path.

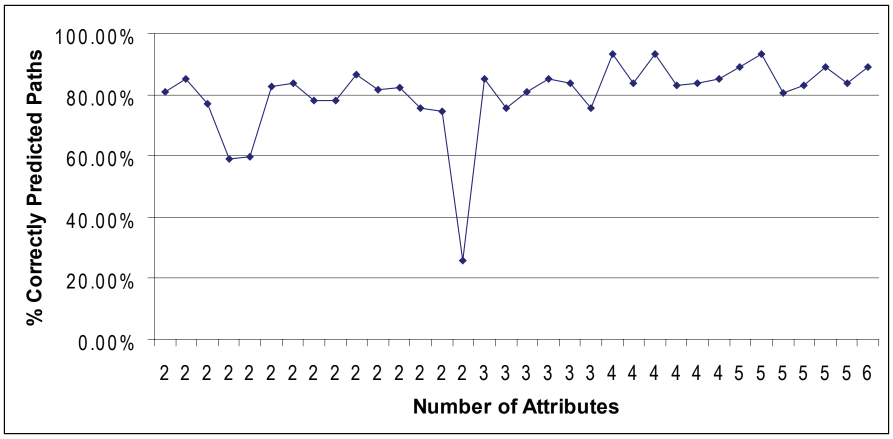
Once profiles are constructed and associated with paths, this data is combined and formatted to be analyzed using Weka (Weka 2004). We have used the J.48 algorithm, which is Weka's implementation of the C4.5 (Hand, Mannila et al. 2001; Weka 2004) decision tree learner to classify profiles.

Each experiment has involved data from 1000 process executions and a variable number of attributes (ranging from 2 attributes to 6 attributes). We have conducted 34 experiments, analyzing a total of 34000 records containing data from process instance executions. Figure 2 shows the results that we have obtained.

Table 3. Additional data stored in the process log

income	loan_type	Loan_ amount	loan_ years	name	SSN	Path
1361.0	Home-Loan	129982.0	33	Bernard-Boar	10015415	FR>CLT>CHL>AHL>NHC>CA
Unknown	Education-Loan	Unknown	Unknown	John-Miller	15572979	FR>CLT>CEL>CA
1475.0	Car-Loan	15002.0	9	Eibe-Frank	10169316	FR>CLT>CCL>ACL>NCC>CA
…	…	…	…	…	…	…

Figure 2. Experimental results



The path mining technique developed has achieved encouraging results. When three or more attributes are involved in the prediction, the system is able to predict correctly the path followed for more than 75% of the process instances. This accuracy improves when four attributes are involved in the prediction, in this case more than 82% of the paths are correctly predicted. When five attributes are involved, we obtain a level of prediction that reaches a high of 93.4%. Involving all the six attributes in the prediction gives excellent results: 88.9% of the paths are correctly predicted. When a small number of attributes are involved in the prediction, the results are not as good. For example, when only two attributes are selected, we obtain predictions that range from 25.9% to 86.7%.

PROCESS MINING

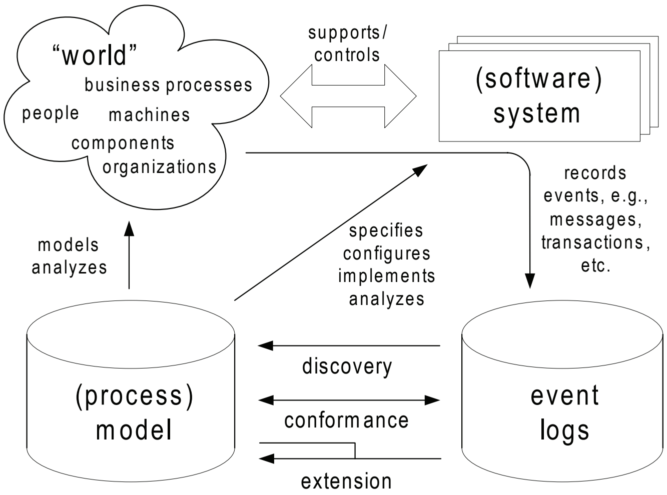
Assuming that we are able to log events, a wide range of process mining techniques comes into reach (Aalst, Dongen, et al. 2003; Aalst, Weijters, Maruster, 2004; Aalst, Reijers, et al. 2007). The basic idea of process mining is to learn from observed executions of a process and (1) to discover new models (e.g., constructing a Petri net that is able to reproduce the observed behavior), (2) to check the conformance of a model by checking whether the modeled behavior matches the observed behavior, and (3) to extend an existing model by projecting information extracted from the logs onto some initial model (e.g., show bottlenecks in a process model by analyzing the event log).

Process discovery: Traditionally, process mining has been focusing on discovery, i.e., deriving information about the original process model, the organizational context, and execution properties from enactment logs. An example of a technique addressing the control flow perspective is the alpha algorithm, which constructs a Petri net model (Aalst, Weijters, Maruster, 2004) describing the behavior observed in the event log. For example, based on a log like the one depicted in Table 1 it is possible to discover the process model shown in Figure 1. In fact to discover this process model, only the columns “Case” and “Task” are needed.

Conformance checking: Conformance checking compares an a-priori model with the observed behavior as recorded in the log. In (Rozinat, Aalst, 2005) it is shown how a process model (e.g., a Petri net) can be evaluated in the context of a log using metrics such as “fitness” (is the observed behavior possible according to the model?) and “appropriateness” (Is the model “typical” for the observed behavior?). In the example, we could compare the observed behavior in Table 1 with the modeled behavior in Figure 1.

Process extension: There are different ways to extend a given process model with additional perspectives

Figure 3. The three types of process mining (discovery, conformance, and extension) and their relations with models and event logs



based on event logs, e.g., decision mining (Rozinat, Aalst, 2006). Starting from a process model, one can analyze how data attributes influence the choices made in the process based on past process executions. A process model can be extended with timing information (e.g., bottleneck analysis). Clearly, decision mining is closely related to path mining. For example, it is possible to highlight bottlenecks in a discovered process model using the timestamps typically present in a log.

Figure 3 shows the three types of process mining introduced above (discovery, conformance, and extension). Each of the techniques involves a model and an event log. In case of discovery the process model is discovered on the basis of the data present in the log. In case of conformance and extension there is already some initial model.

ProM

In recent years ProM (www.processmining.org) has emerged as a broad and powerful process analysis tool, supporting all kinds of analysis related to business processes (such as verification and error checking). In contrast to many other analysis tools the starting point was the analysis of real processes rather than modeled processes, i.e., using *process mining* techniques ProM attempts to extract non-trivial and useful information from so-called "event logs".

Traditionally, most analysis tools focusing on processes are restricted to *model-based analysis*, i.e., a model is used as the starting point of analysis. For example, a purchasing process can be modeled using EPCs (Event-Driven Process Chains) and verification techniques can then be used to check the correctness of the protocol while simulation can be used to estimate performance aspects. Such analysis is *only useful if the model reflects reality*. Therefore, ProM can be used to both analyze models and logs. Basically, ProM supports all types of analysis mentioned before and the goal is to support the entire spectrum shown in Figure 3. It provides a plug-able environment for process mining offering a wide variety of plug-ins for process discovery, conformance checking, model extension, model transformation, etc. ProM is open source and can be downloaded from www.processmining.org. Many of its plug-ins work on Petri nets, e.g., there are several plug-ins to discover Petri nets using techniques ranging from genetic algorithms and heuristics to regions and partial orders. Moreover, Petri nets can be analyzed in various ways using the various analysis plug-ins. However, ProM allows for a wide variety of model types, conversions, and import and export facilities. For example, it is possible to convert Petri nets into BPEL, EPCs and YAWL models and vice versa.

FUTURE TRENDS

In the future, it is expected to see a wider spectrum of applications managing processes in organizations. According to the Aberdeen Group's estimates, spending in the Business Process Management software sector (which includes workflow systems) reached $2.26 billion in 2001 (Cowley 2002).

We are currently extending and improving mining techniques and continue to do so given the many challenges and open problems. For example, we are developing genetic algorithms to improve the mining of noisy logs. Recently, we added a lot of new functionality to the ProM framework. ProM is already able to mine from FLOWer logs and we applied this in a Dutch social security agency. Moreover, we have used ProM to analyze the processes of different hospitals, municipalities, governmental agencies, and banks. We use process mining to analyze the actual use of copiers, medical equipment, and wafer steppers. Furthermore, we are extending ProM with quality metrics to analyze business processes (Vanderfeesten, Cardoso et al. 2007; Mendling, Reijers et al. 2007).

CONCLUSION

Business Process Management Systems, processes, workflows, and workflow systems represent fundamental technological infrastructures that efficiently define, manage, and support business processes. The data generated from the execution and management of processes can be used to discover and extract knowledge about the process executions and structure.

We have shown that one important area of processes to analyze is path mining, i.e. the prediction of the path that will be followed during the execution of a process. From the experiments, we can conclude that classification methods are a good solution to perform path mining on administrative and production processes. We have also shown that business process mining aims at the extraction of knowledge from the behavior observed in the event log. This type of mining can be used to, for example, construct a model based on an event log or to check if reality conforms to the model. Several process mining techniques have been implemented and are made available through the ProM framework.

REFERENCES

Aalst, W.M.P. van der, Dongen, B. F. v., Herbst, J., Maruster, J., Schimm, G. & Reijers, H.A. (2003). Workflow Mining: A Survey of Issues and Approaches. *Data & Knowledge Engineering*, 47(2), 237-267.

Aalst, W.M.P. van der, Reijers, H.A. & Song, M. (2005). Discovering Social Networks from Event Logs. *Computer Supported Cooperative Work*, 14(6), 549–593.

Aalst, W.M.P. van der, Reijers, H. A., Weijters, A.J.M.M., Dongen, B.F. van, Alves de Medeiros, A.K., Song, M. & Verbeek, H.M.W. (2007). Business process mining: An industrial application. *Information Systems,* 32(5), 713-732.

Aalst, W.M.P. van der, Weijters, A.J.M.M. & Maruster, L. (2004). Workflow Mining: Discovering Process Models from Event Logs. *IEEE Transactions on Knowledge and Data Engineering*, 16(9), 1128–1142.

Agrawal, R., Gunopulos, D. & Leymann, F. (1998). Mining Process Models from Workflow Logs. *Sixth International Conference on Extending Database Technology*, Valencia, Spain, Lecture Notes in Computer Science Vol. 1377, Springer, 469-483.

Cardoso, J., Miller, J., Sheth, A., Arnold, J. & Kochut, K. (2004). Modeling Quality of Service for workflows and web service processes. *Web Semantics: Science, Services and Agents on the World Wide Web Journal*, 1(3), 281-308.

Cowley, S. (2002). Study: BPM market primed for growth. Retrieved March 22, 2006, from *http://www.infoworld.com*

Dumas, M., Aalst, W.M.P. van der & A. H. t. Hofstede (2005). *Process Aware Information Systems: Bridging People and Software Through Process Technology*, New York: Wiley-Interscience.

Grigori, D., Casati, F., Dayal, U. & Shan, M. C. (2001, September). Improving Business Process Quality through Exception Understanding, Prediction, and Prevention. *27th VLDB Conference*, Roma, Italy, 2001.

Hand, D. J., Mannila, H. & Smyth, P. (2001). *Principles of Data Mining*. Bradford Book.

Mendling, J., Reijers, H. A. & Cardoso, J. (2007, September). What Makes Process Models Understandable? *Business Process Management 2007*. Brisbane, Australia, 2007, 48-63.

Sayal, M., Casati, F., Dayal, U. & Shan, M. C. (2002, August). Business Process Cockpit. *28th International Conference on Very Large Data Bases*, VLDB'02. Hong Kong, China, 880-883

.Smith, H. and Fingar, P. (2003). *Business Process Management (BPM): The Third Wave*. FL, USA, Meghan-Kiffer Press.

Vanderfeesten, I., Cardoso, J., Mendling, J., Reijers, H. & Aalst, W.M.P. van der. (2007). Quality Metrics for Business Process Models. In L. Fischer (ed.) *Workflow Handbook 2007*. FL, USA: Lighthouse Point, Future Strategies Inc.

Weka (2004). Weka. Retrieved May 12, 2005, from http://www.cs.waikato.ac.nz/ml/weka/.

KEY TERMS

Business Process: A set of one or more linked activities which collectively realize a business objective or goal, normally within the context of an organizational structure.

Business Process Management System: A Business Process Management System (BPMS) provides an organization with the ability to collectively define and model their business processes, deploy these processes as applications that are integrated with their existing software systems, and then provide managers with the visibility to monitor, analyze, control and improve the execution of those processes.

Process Definition: The representation of a business process in a form which supports automated manipulation or enactment by a workflow management system.

Process Log: During the execution of processes, events and messages generated by the enactment system are recorded and stored in a process log. It is an electronic archive in which the history of instances is recorded. It contains various details about each instance, such as starting time, tasks performed and resources allocated.

Task: A task is an "atomic" process: one which is not further subdivided into component processes. It is thus a logical unit of work; in other words a task is either carried out in full or not at all.

Workflow: The automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules.

Workflow Engine: A workflow engine provides the actual management of workflows. Amongst other things, it is concerned with task-assignment generation, resource allocation, activity performance, the launching of applications and the recording of logistical information.

Workflow Management System: A system that defines, creates and manages the execution of workflows through the use of software which is able to interpret the process definition, interact with participants and, where required, invoke the use of tools and applications.