

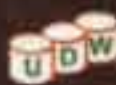
Knowledge Management

&

Intelligent Enterprises

Editors

*Joseph Fong
Daniel Chan
Qing Li
Ronnie Cheung*



Knowledge Management

&

Intelligent Enterprises



Editors

*Joseph Fong, Daniel Chan,
Qing Li & Ronnie Cheung*

This book provides an up-to-date study of the fundamental facets of supporting information technologies: organisation and people, data and information, presentation of information, and process. To improve the effective use of human resources, it is necessary to capture the structure of an organisation and the knowledge of people within the organisation. Data can be processed to produce new information by either inferring hidden implications or summarising a large volume of information into more manageable information units. Information may also be required to be presented using different perspectives so as to meet individual needs and at the same time reinforce policy on data access rights. Routing of information and job tasks is paramount in ensuring that all necessary steps are taken for a given job. These complementary aspects are explored throughout the volume.

The book is a selection of papers presented at the Industrial Session of the 9th IFIP 2.6 Working Conference on Database Semantics. It will be of significant value to researchers, teachers, students and practitioners. The scope embraces design methodology, implementation techniques, and applications development. Specific topics include knowledge management, intelligent enterprises, query transformation, information sharing and retrieval, data warehousing, and knowledge discovery.

کتابخانه سازمان مدیریت توانیر
(صنعت برق)



01BL00005601

World Scientific

www.worldscientific.com

4713 sc

ISBN 981-02-4635-8(pbk)



9 789810 246358

Knowledge Management & Intelligent Enterprises

Industrial Volume
9th IFIP 2.6 Working Conference on
Database Semantics (DS-9) – Semantic
Issues in e-Commerce Systems
organized by the IFIP Working Group 2.6
(Database)

Hong Kong

25–28 April 2001

Editors

Joseph Fong

City University of Hong Kong

Daniel Chan

Systek Information Technology Ltd, Hong Kong

Qing Li

City University of Hong Kong

Ronnie Cheung

Hong Kong Polytechnic University



*Published by World Scientific Publishing Co. Pte. Ltd.
on behalf of the International Federation for Information Processing (IFIP)*

 **World Scientific**
Singapore • New Jersey • London • Hong Kong



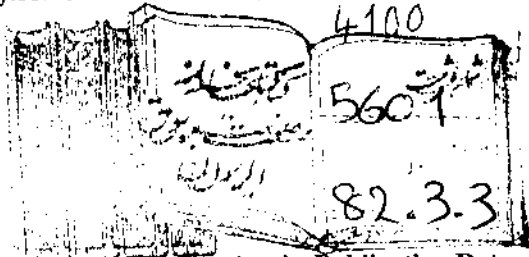
Published by

World Scientific Publishing Co. Pte. Ltd.

P O Box 128, Farrer Road, Singapore 912805

USA office: Suite 1B, 1060 Main Street, River Edge, NJ 07661

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE



British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

Cover: Design by Mr. CHAN Chung Hoi of SYSTEK Information Technology Limited.

KNOWLEDGE MANAGEMENT & INTELLIGENT ENTERPRISES

Copyright © 2001 by World Scientific Publishing Co. Pte. Ltd.

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

ISBN 981-02-4635-8

Printed in Singapore.

Preface

Knowledge Management and Intelligent Enterprises is the Industrial Session in the 9th IFIP 2.6 Working Conference on Database Semantics (DS-9), organised by IFIP Working Group 2.6 (Database) and Hong Kong Web Society. The DS-9 conference is the main core of the trend of Internet Revolution, and is becoming the norm of IT in both academic and industry. The conference is co-hosted by Hong Kong Polytechnic University and the Hong Kong Web Society. As reflected by the diverse submissions from different countries and regions, we expect attendants from both overseas academicians as well as local computer professionals. The event lasts 4 days from April 25 - 28, 2001, in Sheraton hotel in Hong Kong.

The main theme of this industrial track is to improve the effective use of human resources. It is necessary to capture the structure of an organisation and the knowledge of people within the organisation. Data can be processed to produce new information by either inferring hidden implications or summarising large volume of information into more manageable information units. Information may also be required to be presented using different perspectives so as to meet individual needs and at the same time reinforce policy on data access rights. Routing of information and job tasks is paramount in ensuring all necessary steps are taken for a given job. The keynote speech will be presented by Professor John Mylopoulos of University of Toronto on "Managing Knowledge for Business Analysts: The Executive Information Portal".

On behalf of International Federation for Information Processing and Hong Kong Web Society, We would like to take this opportunity to express our thanks to the conference sponsors: ACM Hong Kong Chapter, Hong Kong Polytechnic University, Hong Kong Pei Hua Foundation Limited, British Computer Society (Hong Kong Section) program committee members and organising committee members, who contributed to this function, and trust you will find this a useful and practical conference reference proceedings.

Joseph Fong	Daniel Chan	Qing Li	Ronnie Cheung
<i>Industrial</i>	<i>Industrial</i>	<i>Conference</i>	<i>Organizing</i>
<i>Chair</i>	<i>Chair</i>	<i>Chair</i>	<i>Chair</i>

CONTENTS

Preface	v
PART ONE Knowledge Management	
1. The Executive Information Portal (keynote paper) <i>J. Mylopoulos, A. Barta, R. Jarvis, P. Rodriguez-Gianolli and S. Zhou</i>	3
2. Professional Services Automation: A semantics-based approach for Knowledge Management <i>V. Kashyap, S. Dalal, P. Tukey and C. Behrens</i>	10
PART TWO Data Warehousing	
3. Beyond Data Warehousing – Data Logistics (invited paper) <i>K. Karlapalem</i>	29
4. Architecture of a Data Warehouse with a Database Proxy Server <i>S. M. Huang</i>	34
PART THREE Intelligent Enterprises	
5. Organisation Modelling using LDAP <i>D. K. C. Chan, S. Yuen and S. C. Cheung</i>	53
6. Eguru: A Decision Support System for the assisted design of E-commerce architectures <i>P. Missier, M. Bianchi, A. Zordan and A. Umar</i>	69
7. Office Workflow System Based on the OAR Office Model <i>S. Desai, N. Ambastha, Venkatnarayan V and Binoy E. D.</i>	93

PART FOUR Query Transformation

8. Towards Query Translation from XQL to SQL 113
J. Fong and T. Dillon
9. Rewriting Rules for Semantic Query Transformation in
E-Commerce Applications 130
I. K. Ibrahim, W. Winiwarter and S. Bressan

PART FIVE Information Sharing and Retrieval

10. iDataprovider: An XML-based Mechanism for Internet Data
Exchange Services and Applications 147
E. W. C. Leung and Q. Li
11. Information Retrieval by Semantically Correlated Filamentous
Propagation (CFP) 165
T. Kwok and C. A. Pickover

PART SIX Knowledge Discovery

12. Mining Is-Part-of Association Patterns from Semistructured Data 189
K. Wang and H. Liu
13. WHAT: A Web Hypertext Associated Trail Mining System 205
W. Ng and C. Chan

- Author Index** 221

PART ONE

Knowledge Management

1. The Executive Information Portal: An Extended Abstract¹

John Mylopoulos, Attila Barta, Raoul Jarvis,
Patricia Rodriguez-Gianolli, and Shun Zhou²

Strategic business analysts keep track of trends that are relevant to their organization and its strategic objectives. To accomplish their mission, they monitor news stories and other reports as they become available, looking for evidence that these objectives remain on track, or have encountered obstacles. The paper presents an overview of a prototype enterprise information portal intended to support this type of knowledge work. The system supports three key functions. Firstly, it offers tools for building and analyzing semantic models of strategic objectives, the relationships among them, as well as events and actors who can play a role, positive or negative, in their fulfillment. Secondly, the system uses a powerful query language and the semantic model to assist analysts as they search external sources for relevant material, also provides semi-automatic classification and clustering of documents. Thirdly, documents are placed in an XML format and stored in an XML server. In addition, analysts can annotate or summarize documents and relate them to nodes of the semantic model.

Overview

An important ingredient of the Information Revolution is that individuals and organizations alike have access to orders of magnitude more information than ever before. Moreover, most of this information is available in computer-based forms, including electronic documents, databases and websites. The objective of our research is to develop technologies which help a group of knowledge workers search, find, retrieve, and organize information that is useful to their daily work.

¹ An extended description of this work can be found at <http://www.cs.toronto.edu/~jm/exip.ps>.

² Authors' address: Department of Computer Science, University of Toronto, Toronto, CANADA; {jm,atibarta,prg,jarvis,szhou}@cs.utoronto.ca.

These technologies include data retrieval and analysis for structured data (i.e., database tuples and/or records), semi-structured data (such as hypertext and source code), and unstructured data (such as documents, digitized photos). In general, such technologies are bundled into a solution through an integration architecture often called an *enterprise information portal* (EIP). EIPs are "...*applications that enable companies to unlock internally and externally stored information, and provide users a single gateway to personalized information needed to make informed business decisions*" [Shilakes98]. EIPs are not off-the-shelf software. Rather, they are integrated solutions built out of components. The majority of EIP providers develop their EIP solutions from product suits offered by several cooperating partner companies. Typically, EIPs offer facilities for search, retrieval, analysis and organization of structured data and documents, along with facilities for network connectivity and computing platform interoperation.

Our approach to enterprise information portals is founded on the use of a semantic model that captures knowledge shared by a group of collaborating knowledge workers. This model is used to drive information retrieval, classification, and analysis. For example, consider a group of software engineers who own a legacy software system. An EIP solution for this group would include a semantic model which represents concepts such as (software) global architectures, call graph structures, module interdependencies, versions, maintenance histories, and more. Reverse engineering tools can then extract information from the legacy system and associate it with the relevant components of the semantic model. When a software engineer is working on a particular task, say fixing a bug, she can rely on the EIP (and the semantic model) to find relevant information, such as global variable declarations, who calls the procedure that is being debugged and more. This is precisely the approach explored in [Finnigan97].

We are currently developing a prototype toolset, called the Executive Information Portal (or EXIP), intended to help a group of strategic business analysts working for a large corporation. Their task is to keep track of current events as they unfold, and make sure that their company's strategic objectives remain on track. To work on this task, business analysts scan news stories (Reuter's, CNN, etc.), analysts' reports (Yankee Group, Forrester Research, and the like) and other document sources, looking for relevant material. Once they have decided that a particular document is useful, they add it to their own library, write annotations and prepare memos to be circulated to their colleagues. This work is currently done without any

tool support, or vanilla computer tools (e.g., a web browser and search engine). Our toolset aims to support the semi-automatic search and classification of documents, also the analysis of collected information with respect to a given set of strategic objectives.

The development of our toolset was guided by a number of quality requirements. Firstly, we wanted a system that evolves semi-automatically in the sense that its model and classification scheme may be modified by users, or automatically, thanks to the application of Machine Learning techniques. In addition, the classification of downloaded information is also assumed to evolve, along with the model and the classification scheme. Finally, we expect that our system will be used with minimal feedback from users, who are busy people and can't afford to spend much time training the system's classifiers and other functions.

The remainder of this extended abstract presents the architecture of the EXIP, describes the status of the implementation and notes directions for further research.

System Architecture

The EXIP global architecture is shown in figure 1. The outermost layer of the architecture (bottom part of the figure) includes external information sources, such as CNNfn (cnfn.cnn.com/news/technology/), CBC business news (cbc.ca/business/), the Globe and Mail (globeandmail.com/hubs/rob.html) and Forrester Research, whose reports we assume that the analysts download manually. This layer also includes wrappers for each source, which specify expected outputs for input queries. The information sources may have formats that are structured or semi-structured, proprietary or public domain (e.g., plain text.) In the prototype implementation, we wrap semi-structured (HTML) sources and plain text sources only. Structured sources (such as relational data) are easy to wrap and access, while documents in proprietary formats (PDF, MS-Word, RTF, etc.) can be translated to a common HTML or XML format using commercial tools such as Document Navigator from Verity [Verity].

It is important to note that documents transformed to HTML from a proprietary format are different from "native" HTML documents in that the

former contain only useful information, while the latter contain an abundance of noise (advertising, site navigational menus, etc.) It is the job of the wrapper to locate the information of interest in a page for a given query, and separate it from such noise.

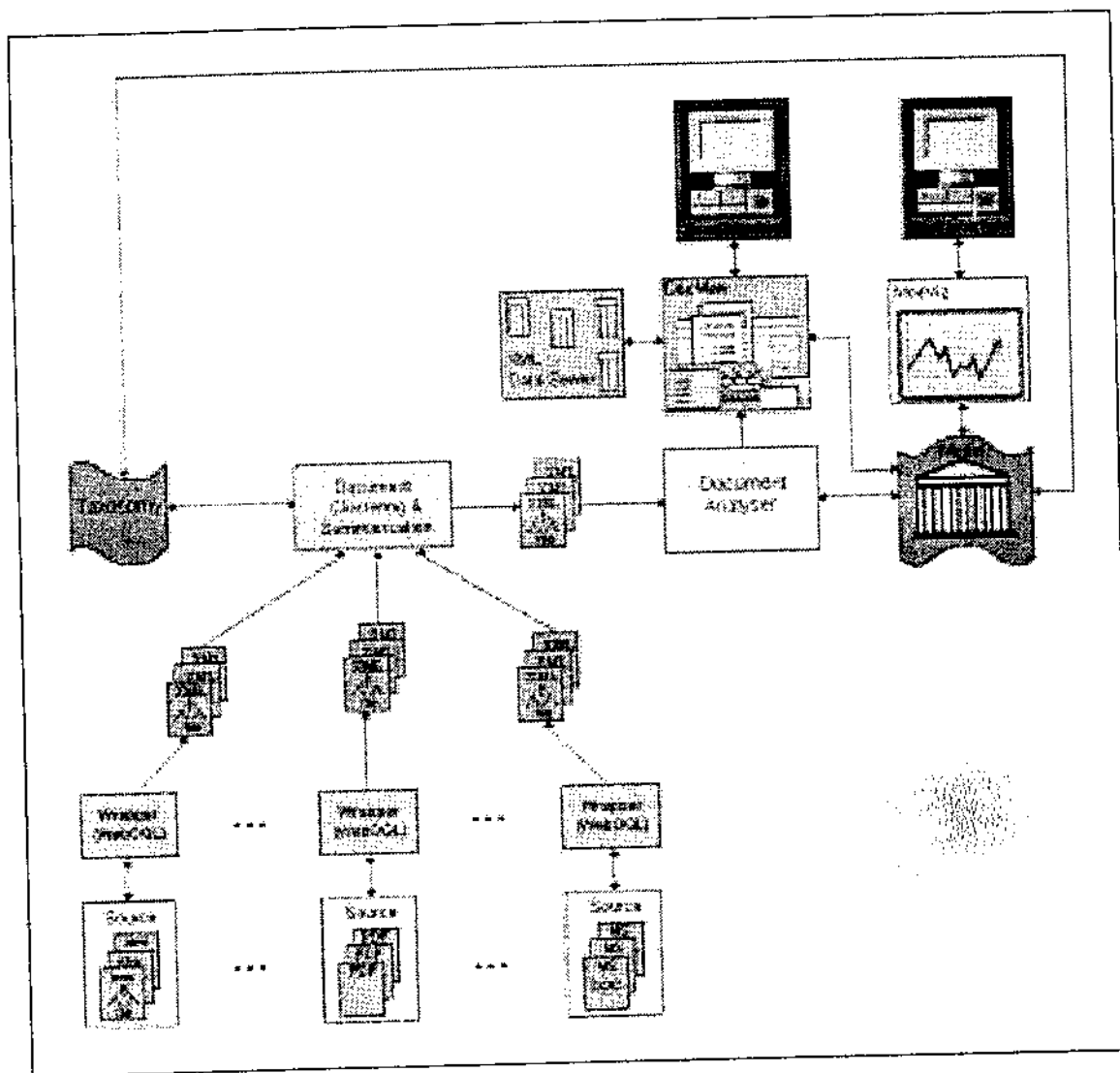


Figure 1: The global architecture of the EXIP.

Another significant difference between the two types of documents is that those in a proprietary format are very likely to be local, easy to access files. In contrast, "native" HTML documents are often buried deep in a website and require considerable navigation in order to be retrieved. To

specify the navigational path that will lead to a document to be retrieved, we use WebOQL, a second-generation web query language [Arocena98].

Wrappers export data in XML format. One important reason for selecting XML over HTML is that XML is fast becoming the lingua franca for information exchange, and its adoption allows us to benefit from a wealth of research in this area. Moreover, XML tags carry more useful information than HTML ones. This information can be used to better index and otherwise process retrieved documents. All exported data are translated into a common document schema, essential for the processing of documents after they have been downloaded.

The Document Clustering and Summarization component uses Machine Learning techniques in order to classify all downloaded documents by relating them to one or more elements of the semantic model. The proposed classification may be approved or overruled by the users of the system, or it may be accepted "as is" when the system is in "automatic" mode. This task is performed through the Document Analysis component. This component was designed so that it can perform reasonably well even when it is shown few training samples.

The semantic model provides a description of the strategic objectives of an organization in terms of goals and subgoals, also the events and actors that can influence any of these goals (positively or negatively). The model can be thought as a network of relationships between goals, actors, events and documents. Thanks to the rich modeling framework, the model supports various forms of analysis. For example, the analyst can visualize if and how the organization is advancing towards achieving its goals, what are the obstacles, critical events to look out for, and what are the dependencies to external organizations.

The shared knowledge of a group of analysts, represented by a semantic model, can be quite large and complex. Moreover, users of the EXIP (that is, business analysts and executives) are usually concerned with strategic planning from different perspectives. To aid the understanding and acquisition of knowledge at different levels of granularity, the user interface of EXIP supports a variety of visualization techniques for browsing/exploring the model as well as the contents of the EXIP.

The Document Management component, DocMan, provides support for document viewing, annotation and manipulation. DocMan manages collections of thousands of documents that have multiple links among them. Furthermore, DocMan supports complex search operations on these

documents, involving both full-text and metadata search. Because the documents and their components (annotations, summaries, links) are in an XML format, DocMan uses an XML Data Server for its operations.

Conclusions

The EXIP is intended to support a group of knowledge workers in retrieving, classifying and using information downloaded from a variety of information sources. Our approach differs from state-of-practice EIPs in that it treats all information managed by the system as semi-structured data, thereby exploiting tools such as declarative query languages and XML data servers. Moreover, our approach adopts a Machine Learning framework for quick learning and continuous evolution of its classifiers and clustering algorithms. Finally, our framework supports lightweight semantic models of relevant domain knowledge which is used for classification, retrieval and analysis.

We are currently completing the implementation and evaluation of the prototype. In addition, we are working on the definition of goal analysis techniques, such as evaluating the satisfiability of top-level goals, given a set of event instances about the application domain. Another form of analysis under review involves the identification of critical goals and events. These are goals or events which have become key to the fulfillment of top-level goals and about which additional information is needed.

Acknowledgements

The research reported in this paper has been funded by the Bell University Laboratories (BUL), the Natural Sciences and Engineering Research Council (NSERC) of Canada, and the University of Toronto. We are grateful to Geoff Riggs (Bell Canada) for introducing us to the world of strategic business analysis, also to Michael Milton and Adele Newton (BUL) for helpful direction and moral support.

References

- [Arocena98] Arocena, G., Mendelzon, A., "WebOQL: Restructuring Documents, Databases and Webs", Proceedings ", International Conference on Data Engineering (ICDE'98), Florida, 1998.
- [Finnigan97] P. Finnigan, I. Kalas, K. Kontogiannis, H. Müller, J. Mylopoulos, S. Perelgut, M. Stanley, K. Wong, "The Software Bookshelf", *IBM Systems Journal*, November 1997.
- [Shilakes98] Shilakes, C., and Tylman, J., "Enterprise Information Portals," Merrill Lynch, November 16, 1998.
- [Verity] <http://www.verity.com/products/docnav/index.html>.