

Filtering the Web to Feed Data Warehouses

Springer-Verlag London Ltd.

Witold Abramowicz, Paweł Kalczyński
and Krzysztof Węcel

Filtering the Web to Feed Data Warehouses



Springer

Witold Abramowicz, MSc, PhD

Paweł, Kalczyński, MSc

Krzysztof Węcel, MSc

Department of Computer Science, The Poznań University of Economics,
al. Niepodległości 10, 60-967 Poznań, Poland

British Library Cataloguing in Publication Data

Abramowicz, Witold

Filtering the Web to feed data warehouses

1.Data Warehousing 2.Information retrieval 3.World Wide Web

I.Title II.Kalczyński, Paweł III. Węcel, Krzysztof

005.7'4

Library of Congress Cataloging-in-Publication Data

Abramowicz, Witold

Filtering the Web to feed data warehouses / Witold Abramowicz, Paweł Kalczyński,
and Krzysztof Węcel.

p. cm.

Includes bibliographical references and index.

ISBN 978-1-4471-1107-8

ISBN 978-1-4471-0137-6 (eBook)

DOI 10.1007/978-1-4471-0137-6

1. Data warehousing. 2. World Wide Web. I. Kalczyński, Paweł, 1977- II. Węcel,
Krzysztof, 1976- III. Title.

QA76.9.D37 A23 2002

658.4'038'0285574--dc21

2002021742

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

<http://www.springer.co.uk>

© Springer-Verlag London

Originally published by Springer-Verlag London Limited in 2002

The use of registered names, trademarks etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Preface

Every human activity, particularly business activity, is based on some information resources that ought to be collected beforehand. Business people collect information in order to make effective, that is, fast and accurate, decisions. The constantly growing competition on markets and the dynamic nature of present-day business enhances the demand for faster and better decisions. Such decisions first and foremost depend on the quality of information resources and the organizational ability to utilize them.

Most of the contemporary organizations, regardless of their size, have learned to collect their internal data and to transform it into useful information to support the decision-making process. However, a posteriori information derived from organizational data is insufficient to support decision-making processes in terms of today's competitiveness. Organizations realize that fact and make serious efforts to assure persistent supply of (or access to) external information to their resources.

In recent years, the Internet has become an almost unlimited source of information (including business information), and since then several high-quality-business-content providers such as Money CNN¹, Financial Times² or A.C. Nielsen³ have emerged on the market.

Today, business content providers on the Web maintain an immeasurable amount of information resources, and the real problem is finding (filtering) useful pieces of information in reasonable time. This problem results from the characteristic features of the contemporary Internet as well as from user inability to deal with information sources on the Web. What is more, the acquired information must be incorporated in the organizational information resources and there is a lack of methodologies and tools.

As the result, two different worlds exist in businesses today: the world of internal information and the world of external information. The lack of true connections between these separate worlds negatively affects the effectiveness of the decision-making process for apparent reasons. Business people are forced to use different interfaces to information resources. This sometimes makes organizations decide not to utilize the Web at all and to make use of traditional sources of business information such as business newspapers.

This book provides a new idea of bringing the world of data and the world of documents together in the organizational information system. The concept of filtering the Web to feed data warehouses is based on the existing knowledge in the

¹ <http://money.cnn.com/>

² <http://ft.com/>

³ <http://www.acnielsen.com/>

area of data and document management. The idea is illustrated by the working implementation of the described solution and by some experiment results.

Intended Audience

This book addresses several aspects of data management, document management and temporal information. It is first and foremost intended for practitioners, who deal with information management in organizations, and researchers who develop solutions in this area. Software engineers may be interested in the sophisticated data and information structures presented in this book. Lecturers may also use this book as a supplement in advanced courses on information management in business organizations.

Poznań, Poland
February 2002

*Witold Abramowicz
Paweł Jan Kalczyński
Krzysztof Węcel*

Table of Contents

CHAPTER 1 INTRODUCTION	1
1.1 Information Systems	1
1.2 Information Filtering Systems.....	2
1.3 Database Systems.....	2
1.3.1 Transactional Systems	3
1.3.2 Analytical Systems	4
1.4 Organization of this Book	5
CHAPTER 2 DATA WAREHOUSE: CORPORATE KNOWLEDGE REPOSITORY	7
2.1 Introduction.....	7
2.2 Data Warehouse Definition and Features.....	7
2.2.1 Definition.....	7
2.2.2 Metadata	8
2.2.3 Characteristic Features of Data in the Data Warehouse	9
2.3 Data Warehouse System	11
2.3.1 Architecture of the Data Warehouse System	11
2.3.2 Metadata Structures	16
2.3.3 Data Warehouse Products.....	22
2.4 Deploying Data Warehouse in the Organization.....	23
2.4.1 Data Warehouse Life Cycle.....	23
2.4.2 Analysis and Research.....	24
2.4.3 Identifying Architecture and Demands.....	24
2.4.4 Design and Development.....	25
2.4.5 Implementation and On-going Administration.....	25
2.5 Knowledge Management in Data Warehouses	26
2.5.1 Knowledge Management	26
2.5.2 Knowledge in Terms of Data Warehousing.....	27
2.5.3 Knowledge Discovery in Data Warehouses	28
2.5.4 Significance of Business Metadata	29
2.6 Evolution of the Data Warehouse	29
2.6.1 Criticism of the Traditional Data Warehouse	29
2.6.2 Virtual Data Warehouse	30
2.6.3 Information Data Superstore.....	30
2.6.4 Exploration Warehouse	30
2.6.5 Internet/Intranet Data Warehouse	32
2.6.6 Web Farming	33
2.6.7 Enterprise Information Portals.....	35

2.7	Chapter Summary	36
2.8	References	37
CHAPTER 3 KNOWLEDGE REPRESENTATION STANDARDS		41
3.1	Introduction	41
3.1.1	Basic Concepts	41
3.1.2	Metadata Representation	42
3.1.3	Metadata Interoperability.....	43
3.1.4	Theory of Metadata	43
3.2	Markup Languages	45
3.2.1	Background.....	45
3.2.2	XML Document.....	46
3.2.3	Document Presentation	47
3.2.4	Document Linking	47
3.2.5	Programming Interfaces.....	47
3.3	Dublin Core.....	48
3.3.1	Dublin Core Metadata Elements	48
3.3.2	Dublin Core in HTML	49
3.4	Warwick Framework.....	49
3.5	Meta Content Framework	51
3.5.1	Origins of MCF	51
3.5.2	Conceptual Building Blocks of MCF	51
3.5.3	XML Syntax	52
3.5.4	Directed Labelled Graph Formalism	54
3.6	Resource Description Framework	55
3.6.1	Background.....	55
3.6.2	Formal RDF Data Model.....	56
3.6.3	The RDF Syntax	58
3.6.4	RDF Schema.....	62
3.7	Common Warehouse Metamodel.....	65
3.7.1	History of OMG Projects.....	65
3.7.2	Objectives of the CWM.....	66
3.7.3	Metadata Architecture	66
3.7.4	CWM Elements	69
3.7.5	Conclusions for CWM.....	70
3.8	Chapter Summary	71
3.9	References.....	72
CHAPTER 4 INFORMATION FILTERING AND RETRIEVAL FROM WEB SOURCES		75
4.1	Introduction.....	75
4.1.1	Document, Information, Knowledge	75
4.1.2	Indexing.....	76
4.1.3	Hypertext.....	77
4.1.4	Information on the Web.....	77
4.1.5	Constraints of this Book	78

4.2	Information Retrieval Systems.....	79
4.2.1	Definitions	79
4.2.2	Information Retrieval System Architectures and Models.....	81
4.2.3	Sample Information Retrieval Systems	87
4.3	Information Filtering Systems.....	88
4.3.1	Filtering Versus Retrieval.....	88
4.3.2	Information Filtering Models and Architectures	89
4.3.3	Sample Filtering Systems	92
4.4	Internet Sources of Business Information	93
4.4.1	Business View on Internet Information Sources	93
4.4.2	General Characteristics of Business Information Sources.....	94
4.4.3	Information Overflow.....	95
4.5	Filtering the Web to Feed Business Information Systems	97
4.5.1	Problems with Web Filtering and Retrieval.....	97
4.5.2	New Information Filtering System Model Proposal.....	98
4.5.3	Transparent Filtering and Retrieval	100
4.6	Chapter Summary	101
4.7	References.....	101
CHAPTER 5 ENHANCED DATA WAREHOUSE.....		105
5.1	Introduction.....	105
5.2	Justification of the Need for Integration	106
5.2.1	Value of Knowledge.....	106
5.2.2	Attention Economy.....	107
5.2.3	Content Management and Lifecycle of Content	108
5.2.4	Example of Integration: Metadata and Data	110
5.3	Preliminary Vision of the System	110
5.3.1	Analytical Point of View	111
5.3.2	Trends.....	111
5.3.3	Goals of the System.....	111
5.3.4	User Requirements Towards the Information Retrieval Systems..	112
5.4	Software Agents.....	113
5.4.1	Introduction	113
5.4.2	Intelligent Agents or Just Agents?.....	113
5.4.3	Software Agents or Just Agents?.....	113
5.4.4	Possible Applications of Agents	114
5.4.5	Definitions of Software Agents	115
5.4.6	Agent Properties	117
5.4.7	Classifications of Software Agents.....	118
5.4.8	Agent-based Systems and Multi-agent Systems	120
5.5	Proposed Solution: enhanced Data Warehouse.....	121
5.5.1	Introduction	121
5.5.2	Overview of the eDW System	122
5.5.3	Assumptions for the eDW System.....	124
5.5.4	Components.....	126
5.5.5	Agent-based System Architecture	127

5.5.6	Logging Server	128
5.5.7	Profiling Server.....	128
5.5.8	Source Agent Server	129
5.5.9	Document Server	129
5.5.10	Properties of eDW Agents	130
5.6	Formal Model of eDW	131
5.6.1	CSL: The Extension of the Organizational Metamodel.....	131
5.6.2	Time Consistency among Documents and Warehouse Data	136
5.6.3	DWL: The Intranet Collection of Relevant Documents for the Data Warehouse	139
5.6.4	enhanced Data Warehouse Report: The Final Product of the eDW System	141
5.6.5	Formal Definitions of eDW Agents.....	144
5.7	System Implementation.....	146
5.7.1	Programming Environment	146
5.7.2	System Control Centre.....	147
5.7.3	Communication	148
5.7.4	Status	148
5.7.5	Configuration File.....	149
5.7.6	Logging Server	150
5.8	Chapter Summary	151
5.9	References.....	152
CHAPTER 6 PROFILING		155
6.1	Introduction.....	155
6.2	Personalization and Data Warehouse Profiles.....	155
6.2.1	Classification of Information	155
6.2.2	Personalization.....	156
6.2.3	Personalization in Data Warehouses and its Aspects.....	156
6.2.4	Overview of Profile Creation.....	157
6.2.5	Data Warehouse Profiles	159
6.3	Algorithms Specification	161
6.3.1	Algorithm for Creating Warehouse Profiles	161
6.3.2	Computational Complexity.....	164
6.3.3	Thesauri	165
6.4	Profiling Server.....	166
6.4.1	Basic Assumptions	166
6.4.2	Profiling Agent	166
6.4.3	User Interface in Profiling Application	168
6.4.4	Sample Results	170

6.5 Chapter Summary175

6.6 References.....175

CHAPTER 7 SOURCE EXPLOITATION.....179

7.1 Introduction.....179

7.2 Sample Business Content Providers.....179

7.2.1 Sample Business Gateways179

7.2.2 Sample Business Search Engines181

7.2.3 Sample Business Portals and Vortals.....181

7.2.4 Sample Business Online Databases184

7.3 Information Ants to Filter Information from Internet Sources.....186

7.3.1 Introduction186

7.3.2 Ant Colony Optimization186

7.3.3 Environment for Information Ants187

7.3.4 Information Ants to Filter Information from the Web.....189

7.3.5 Experiment with Ant-like Navigation.....190

7.3.6 Advantages and Drawbacks of the Proposed Solution192

7.4 Indexing Parser193

7.4.1 Parsing Web Documents.....193

7.4.2 Indexing Web Documents197

7.5 Transparent Filtering in the eDW System.....198

7.5.1 Building Warehouse Profiles.....198

7.5.2 Registering Sources199

7.5.3 Source Exploration200

7.5.4 Source Penetration.....201

7.6 Chapter Summary201

7.7 References.....202

CHAPTER 8 BUILDING DATA WAREHOUSE LIBRARY203

8.1 Introduction.....203

8.1.1 Characteristics of WWW: A Dream of Non-volatile Internet203

8.1.2 Digital Libraries.....204

8.2 Time Indexing.....205

8.2.1 Finite State Automaton205

8.2.2 Time Indexer207

8.2.3 Trapezoidal Time Indices212

8.2.4 Simple Overlap Measure for Trapezoidal Time Indices212

8.3 Experiment with Time Indexing214

8.3.1 Experiment with Time Indexing Real-World Documents214

8.3.2 Conclusions for the eDW System.....223

8.4 Future Trends: Multimedia Indexing225

8.4.1 Introduction225

8.4.2 Filtering Web Documents.....225

8.4.3 Neural Nets for Image Categorization225

8.4.4 The Proposed Solution – Perceptron Categorization Tree.....226

8.4.5 Advantages and Drawbacks.....228

- 8.4.6 Application for eDW228
- 8.5 Chapter Summary229
- 8.6 References.....229
- CHAPTER 9 CONTEXT QUERIES AND ENHANCED REPORTS231**
 - 9.1 Introduction.....231
 - 9.2 Context Queries.....231
 - 9.2.1 Definition of Context.....231
 - 9.2.2 Justification of Transparent Retrieval.....232
 - 9.2.3 Elements of Context232
 - 9.2.4 Conceptual Similarity Measure233
 - 9.2.5 Simple Temporal Similarity Measure.....233
 - 9.2.6 Parameterized Temporal Similarity Measure234
 - 9.2.7 Pertinence235
 - 9.3 enhanced Report.....235
 - 9.3.1 User Interface in Accessing the Information236
 - 9.3.2 How enhanced Report is Created.....236
 - 9.4 Reporting Application.....237
 - 9.4.1 Basic Assumptions237
 - 9.4.2 Description of the Algorithms239
 - 9.4.3 Context Query Agent.....241
 - 9.4.4 Computational Complexity.....242
 - 9.4.5 User Interface in Reporting Application.....246
 - 9.4.6 Results248
 - 9.5 Histograms: The Helpful Tool for Analysis.....252
 - 9.5.1 Non-parameterized Histogram.....253
 - 9.5.2 Past-oriented Analysis253
 - 9.5.3 Future-oriented Analysis254
 - 9.5.4 General Documents255
 - 9.5.5 Detailed Documents255
 - 9.5.6 Compact and Dispersed Histograms.....256
 - 9.6 Chapter Summary258
 - 9.7 References.....258
- CHAPTER 10 CONCLUSIONS.....261**
 - 10.1 Concluding Remarks.....261
 - 10.2 Improvements262
 - 10.3 Open Issues and Future Work.....262
- INDEX..... 265**