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. Niepodleglosci 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.Kalczynski, Pawel III. Wecel, Krzysztof
005.7'4

Library of Congress Cataloging-in-Publication Data
Abramowicz, Witold
Filtering the Web to feed data warehouses / Witold Abramowicz, Pawel Kalczynski, and Krzysztof Wecel.
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. Kalczynski, Pawel, 1977- II. Wecel, Krzysztof, 1976- III. Title.
QA76.9.D37 A23 2002
658.4'038'0285574--dc21 2002021742

http://www.springer.co.uk

Originally published by Springer-Verlag London Limited in 2002

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 CNN[1], Financial Times[2] or A.C. Nielsen[3] 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