2015

Web mining for the integration of data mining with business intelligence in web-based decision support systems

Integration of Data Mining in Business Intelligence Systems

Ana Azevedo
Algoritmi R&D Center/University of Minho, Portugal & Polytechnic Institute of Porto/ISCAP, Portugal

Manuel Filipe Santos
Algoritmi R&D Center/University of Minho, Portugal

A volume in the Advances in Business Strategy and Competitive Advantage (ABSCA) Book Series

BUSINESS SCIENCE Reference
An Imprint of IGI Global

Managing Director: Lindsay Johnston
Production Editor: Christina Henning
Development Editor: Allison McGinniss
Acquisitions Editor: Kayla Wolfe
Typesetter: John Crodian
Cover Design: Jason Mull

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

Library of Congress Cataloging-in-Publication Data

Integration of data mining in business intelligence systems / Ana Azevedo and Manuel Filipe Santos, editors.
pages cm
Includes bibliographical references and index. Summary: “This book investigates the incorporation of data mining into business technologies used in the decision making process, emphasizing cutting-edge research and relevant concepts in data discovery and analysis”-- Provided by publisher. ISBN 978-1-4666-6477-7 (hardcover : alk. paper) -- ISBN 978-1-4666-6478-4 (ebook) -- ISBN 978-1-4666-6480-7 (print & perpetual access) 1. Business intelligence. 2. Data mining. I. Azevedo, Ana, editor. II. Santos, Manuel Filipe, editor.

HD38.7.I5438 2015
658.4’03802856312--dc23

This book is published in the IGI Global book series Advances in Business Strategy and Competitive Advantage (ABSCA) (ISSN: 2327-3429; eISSN: 2327-3437)

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

Chapter 7
Web Mining for the Integration of Data Mining with Business Intelligence in Web-Based Decision Support Systems

Marcos Aurélio Domingues
University of São Paulo, Brazil

Alípio Mário Jorge
University of Porto, Portugal

Carlos Soares
University of Porto, Portugal

Solange Oliveira Rezende
University of São Paulo, Brazil

ABSTRACT

Web mining can be defined as the use of data mining techniques to automatically discover and extract information from web documents and services. A decision support system is a computer-based information sy Analysis stem that supports business or organizational decision-making activities. Data mining and business intelligence techniques can be integrated in order to develop more advanced decision support systems. In this chapter, the authors propose to use web mining as a process to develop advanced decision support systems in order to support the management activities of a website. They describe the Web mining process as a sequence of steps for the development of advanced decision support systems. By following such a sequence, the authors can develop advanced decision support systems, which integrate data mining with business intelligence, for websites.

DOI: 10.4018/978-1-4666-6477-7.ch007

INTRODUCTION

The management of web sites imposes a constant demand for new information and timely updates due to the increase of services and content that site owners wish to make available to their users, which in turn is motivated by the complexity and diversity of needs and behaviors of the users. Such constant labor intensive effort implies very high financial and personnel costs.

A decision support system is a computer-based information system that supports business and organizational decision-making activities. The integration of data mining with business intelligence techniques is important for the development of advanced decision support systems. Data mining is the application of specific algorithms for extracting patterns from data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Business intelligence can be presented as a technology to collect and store data, analyze it using analytical tools, and deliver information and/or knowledge, facilitating reporting, querying, and allowing organizations to improve decision making (Azevedo, & Santos, 2011). An advanced decision support system, which integrates data mining with business intelligence, can support several management activities of a web site, reducing the effort to manage it.

Web mining is usually defined as the use of data mining techniques to automatically discover and extract information from web documents and services (Etzioni, 1996). According to (Kosala, & Blockeel, 2000), web mining is commonly categorized into three areas: web content mining that describes the discovery of useful information from content (Chakrabarti, 2000), web structure mining that analyzes the topology of web sites (Chakrabarti, Dom, Kumar, Raghavan, Rajagopalan, Tomkins, Gibson, & Kleinberg, 1999), and web usage mining that tries to make sense of the data generated by the navigation behavior and user profile (Srivastava, Cooley, Deshpande, & Tan, 2000). These areas naturally overlap and complement one another.

In this chapter, we propose to use web mining as a process to develop advanced decision support systems in order to support the management activities of a web site. We describe the web mining process as a sequence of steps for the development of advanced decision support systems. By following such a sequence, we can develop advanced decision support systems which integrate data mining with business intelligence. We have applied our proposed web mining process in the development of intelligent monitoring/management systems to guarantee the quality of web sites. Examples of monitoring activities include:

- **Usage:** Keep track of the paths users take during their accesses, the efficiency of pages/hyperlinks in guiding the users to accomplish their goals;
- **Users:** How users are grouped taking into account their browsing behavior, how groups change with time, how groups of users relate with the success of the site;
- **Data quality:** How adequate the content and meta-data of a web site are;
- **Automation:** The effect of personalization actions. For instance, if users are following the recommendations of products and pages or not.

This chapter is organized as follows: We start by presenting some work related to web mining. Then, we present our proposal. We show the main data sources that are used in the web mining process. We pay special attention to the preprocessing and storage of web data for use in web site management. Then, we present a case study where web mining is used as a process to develop an advanced decision support system that monitors the quality of the meta-data describing content in an e-news web portal, supporting decision making (Domingues, Soares, & Jorge, 2013). Finally, we present future research directions and conclusion.

RELATED WORK

In this work, we have used web mining to develop advanced decision support systems. However, the web mining process can be used to address different problems. In this section, we briefly describe other uses of web mining introduced by the web mining research community:

- **Web Personalization/Recommendation:** The user navigation behavior can be used to personalize web pages by making dynamic recommendations (e.g., pages, services, etc) for each web user (Anand, & Mobasher, 2003);
- **Categorization/Clustering of Content:** Content data can be used to categorize/cluster web pages into topic directories (Chakrabarti, 2000);
- **Automatic Summarization of Content:** The goal is to construct automatically summaries from web page text content (Zhang, Zincir-Heywood, & Milios, 2004). An example of such application is the presentation of summaries by search engines;
- **Extraction of Keywords from Web Pages:** A keyword is a word or a set of words which characterizes the content of a web page or site, and is used by users in their search process. Using content and usage information from a web page/site, we can extract/identify keywords which attract and retain users (Velasquez, & Palade, 2008);
- **Web Page Ranking:** Hyperlinks can be used to rank web pages, in accordance with the interest of the user, such as in search engines (Page, Brin, Motwani, & Winograd, 1999);
- **Web Caching Improvement:** Access patterns extracted from web logs can be used to extend caching policies in order to improve the performance of web accesses (Bonchi, Giannotti, Gozzi, Manco, Nanni, Pedreschi, Renso, & Ruggieri, 2001);
- **Clickstream and Web Log Analysis:** The access logs can also be used to perform other types of analyses, from simple access statistics to user behavioral patterns, that help to improve the quality of web sites (Spiliopoulou, & Pohle, 2001; Joshi, Joshi, & Yesha, 2003);
- **Analysis of Web Site Topology:** Web content and hyperlinks are used to analyze the topology of a web site and improve its organization, possibly reducing the number of alternative pages/hyperlinks that must be considered when we browse a web site (Wookey, & Geller, 2004);
- **Identifying Hubs and Authorities:** Hyperlinks can also be used to identify hubs (directory pages) and authorities (popular pages) (Chakrabarti, Dom, Kumar, Raghavan, Rajagopalan, Tomkins, Gibson, & Kleinberg, 1999). A hub is a page that points to many other pages. An authority is a page that is pointed to by many different hubs;
- **Identifying Web Communities:** Hubs and Authorities can be combined to identify web communities, which are groups of pages sharing the same subject (Kumar, Raghavan, Rajagopalan, & Tomkins, 1999);
- **OLAP Analysis:** The historical evolution of web data (e.g., usage, content and structure data) is analyzed on several perspectives/dimensions (Hu, & Cercone, 2004).

OUR PROPOSAL

Our proposal consists of using web mining as a process to develop advanced decision support systems in order to support the management activities of a web site. In the next sections, we describe the web mining process as a sequence of steps for

the development of advanced decision support systems: defining web data, pre-processing web data, web data warehousing, and defining pattern discovery and analysis. Then, we present a case study that uses our proposal for developing an advanced decision support system to manage the quality of a web site.

DEFINING WEB DATA

In web mining, data can be collected at the server-side, client-side, proxy server and/or obtained from an organization's database (business or consolidated web data). Different types of data can be used in web mining and, consequently, in web site management (Srivastava, Cooley, Deshpande, & Tan, 2000):

- **Content:** The actual data in web pages. These usually consist of structured and unstructured textual content as well as other multimedia content;
- **Structure:** Data that describe the organization of the pages. These include intra-page structure information (the layout of various HTML or XHTML tags within a given page) and inter-page structure information (the hyperlinks connecting one page to another page);
- **Usage:** Data that describe the usage of web pages (accesses), such as IP addresses, page references and date;
- **User Profile:** Data that provide information about the users of the web site. These include data from registration and customer/user profile.

In this work, we focus on usage, content and structure data. These are used as input for the most common web site management applications (Velasquez, & Palade, 2008). However, there are other web data which can be collected and used to manage a web site. For example, Claypool, Brown, Le, & Waseda (2001) designed a web browser to collect information about the user's behavior regarding mouse, scrollbar and keyboard activities on a web page. So, for the development of a decision support system, by using our proposal, we should first define the web data that will be used as input for the system.

PRE-PROCESSING WEB DATA

Content, structure and usage data must be pre-processed and stored in a database to facilitate their use by data mining techniques. In this section, we discuss the pre-processing of such data.

Usage Data

According to Peterson (2004), usage data can be obtained from web access logs and/or page tagging. Tagging consists in pieces of code placed on a page to notify when the page is accessed. Here, we focus on web access logs as usage data.

The pre-processing of web logs is possibly the most difficult task in the pre-processing of web data due to the incompleteness of the available data. Two examples of causes of web access logs incompleteness/fail (i.e., to represent all interaction between users and the site) are local caching and proxy servers (Cooley, Mobasher, & Srivastava, 1999). With local caching, all requested pages are cached and when a user hits such pages, the cached versions are displayed and the web access log does not register the repeated access. On the other hand, using a proxy server, all requests are handled by such a server, therefore, the requests will have the same identifier (i.e., proxy identifier), even though they represent different users. Thus, as a web access log may not contain a complete interaction of users with the site, a special attention to its pre-processing must be paid.

Various web servers generate different formats of logs. The most common is the Common Log Format (Kimball, & Merz, 2000). In Table 1, we

Table 1. Data files in a web server log

Data Field	Description
Host	Domain name of the client computer or its IP address if the name is not available.
Ident	Identity information of the client computer in a particular TCP connection. If the value is not present, it is indicated by a character "-".
Authuser	User identification used in the request of a password protected page/file. The character "-" indicates that the value is not present.
Time	Date and time that an access/request reaches the server. A common format for these data is [day/month/year:hour:minute:second zone].
Method	The method to access a page/file. The most common are GET and POST. GET encodes data into the URI. POST places the data as part of the request (i.e., in the body of the request). According to (Berners-Lee, & Connolly, 1995), GET should be only used to retrieve the content of the web page to the user, while POST should not be only used to retrieve the web content, but also to post information to the web server.
Request	The URI of a page/file requested by the browser. An URI (Uniform Resource Identifier) is a compact string of characters for identifying a web resource (Berners-Lee, Fielding, & Masinter, 1998).
Protocol	The version of the protocol used by the browser to retrieve pages/files (HTTP/1.0 and HTTP/1.1 are the most common values).
Status	A three digit code which shows the request status. For example, 200 means that the page/file was retrieved successfully, 404 means that the page/file was not found, and so forth.
Bytes	Number of bytes transferred from the web server to the client computer.
Referrer	A text string with the URI of the hyperlink that was followed to the current page.
User-agent	The identification of operating system and browser software used by the user.

present the data fields which are commonly stored in logs. Figure 1 shows an example of the Common Log Format. Each line in the log represents one access/request to the web site. For example, line 1 of Figure 1 indicates that the web page index.html was accessed by IP address 127.0.0.1 at October 10, 2000 at 13 hours and 55 minutes and 16 seconds, using the protocol GET, and the transmission of the web page, with 567 bytes, occurred with success (Status 200).

Nowadays, an organization may use the load balancing technique to distribute the workload of its web site across several web servers to optimize resources, maximize throughput, minimize response time and avoid overload. In this case, the web access logs are also distributed across several web servers and must be put together into a joint log file before being pre-processed. In (Tanasa, & Trousse, 2004; Raju, & Satyanarayana, 2008), the authors propose to merge the log files from the several web servers and sort the requests/accesses in ascending order based on access time, taking into account the synchronization of web server clocks and the time zone differences.

After merging the log files, we have to clean the joint web access log. According to (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), data cleaning involves various activities to produce data which can be exploited successfully. Here, these activities consist of removing the irrelevant requests,

Figure 1. Example of the common log format

```
## Host Ident Authuser [Time] "Method Request Protocol" Status Bytes
127.0.0.1 - - [10/Oct/2000:13:55:16 -0700] "GET /index.html HTTP/1.0" 200 567
127.0.0.1 - - [10/Oct/2000:13:55:20 -0700] "GET /header.gif HTTP/1.0" 200 2326
127.0.0.1 - - [10/Oct/2000:13:55:23 -0700] "GET /logo.gif HTTP/1.0" 200 1316
```

as for example, images and scripts; removing the data fields/columns that are not essential from the analytical point of view; and removing the requests from robots (also called spiders), which are automatically generated by software tools and do not represent human browsing behavior (Cooley, Mobasher, & Srivastava, 1999; Pabarskaite, & Raudys, 2007).

The next two usage data pre-processing activities are the identification of users and sessions. A user is a person who browses the web or enters a web site. A session is a sequence of pages accessed during a single visit to a web site (Cooley, Mobasher, & Srivastava, 1999). We use cookies to identify users and sessions. Cookies enable a web server to indirectly store a text string in a client computer. This text can be later retrieved by the same server (Kimball, & Merz, 2000). This mechanism enables the identification of users returning to a web site, therefore, it can be used to identify users and sessions. If cookies are not present, we can use some heuristics to determine users and sessions. In (Pabarskaite, & Raudys, 2007) the following heuristics are employed for user identification:

- By the http login, which consists in using information entered by users themselves using the http login system. The information is maintained in the Ident and Authuser data fields of the log file;
- By the client type, using the assumption that two requests with the same IP address but different User-agent information are made by two different users (Pirolli, Pitkow, & Rao, 1996);
- Using site topology, where it is assumed that a request is made by a new user if the page is not accessible from the previous set of pages (Cooley, Mobasher, & Srivastava, 1999).

With respect to session identification, in (Pabarskaite, & Raudys, 2007) some heuristics are presented:

- **Time Gap:** When the time gap between two page requests, made by the same user, exceeds some threshold, the pages are assumed to be part of different sessions (Cooley, Mobasher, & Srivastava, 1999). The most common threshold is 30 minutes;
- **Maximum Forward References:** A new session is started after the first repetition of a page in a sequence (Chen, Park, & Yu, 1996);
- **Referrer:** Given two consecutive pages i_1 and i_2, a new session is started if i_1 is not the referrer of i_2 (Berendt, Mobasher, Nakagawa, & Spiliopoulou, 2002).

The last usage data pre-processing activity consists of selecting the data fields from the log file and loading them into a database in order to facilitate the use of web data by data mining algorithms.

Content Data

Web pages consist of structured and unstructured textual content as well as other multimedia content. Usually, the content of a page is written in HTML or XHTML code. An example of a web page and its HTML representation is presented in Figure 2.

Before pre-processing the content of a page, we have to create a local version of it. For that, we can use a crawler, which is a program that visits web sites and reads their pages in a methodical and automated manner. Then, after creating the local version of the web page, the content pre-processing consists of cleaning up the HTML code (to obtain a standard compliant code that facilitates its manipulation), selecting free text and

Figure 2. Example of a web page with its HTML code on the left side

performing searches for tags containing elements which are of interest (e.g., titles, tables and etc). For example, we can parse a web page and search for the tag **title** to retrieve the title of this page.

The content of static web pages can be easily pre-processed by cleaning up and parsing the HTML code. On the other hand, dynamic web pages present a harder challenge. For example, content management systems which draw upon databases to construct the pages may be able of forming more pages than can be practically pre-processed. If only a subset of pages of a web site is pre-processed, the automation of the site may be skewed.

Structure Data

A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an *intra-document hyperlink*, and a hyperlink that connects two different pages is called an *inter-document hyperlink*. Thus, hyperlinks connect web pages, and can be represented as pairs of pages.

Here, pre-processing consists in performing searches for the tags *href* to identify the hyperlinks, and consequently, the structure of the site. An example of a hyperlink is presented in Figure 2 (highlighted ellipses). The structure pre-processing also poses some challenges, as for example, the identification of hyperlinks when they are embedded in web pages developed using technologies like Adobe Flash[1], which allows the creation of interactive content.

After being pre-processed, the web data are loaded into a database to facilitate their use by data mining techniques.

WEB DATA WAREHOUSING

A database to store rich web data is an essential component for a web site management system. A traditional database system is not adequate

for this purpose, because this type of system is developed for transactional tasks, and web site management is essentially an analytical task. Thus, a more adequate database system for supporting web site management is a data warehouse, which is developed for analytical tasks (Chaudhuri, & Dayal, 1997).

A data warehouse can be defined as *a subject oriented, integrated, time-variant, and non-volatile collection of data in support of management decision making process* (Inmon, 2005) or as *a copy of transactional data specifically structured for queries and analyses* (Kimball, & Ross, 2002).

Kimball, & Merz (2000) refer to the application of a data warehouse to web data. The data warehouse, or simply Webhouse, is seen as a solution to store everything related to the clickstream in a web site. However, more recent works, such as (Velasquez & Palade, 2008), also see the data warehouse as a solution to store data related to the content and structure of a web site besides its clickstream.

The design of a data warehouse starts with the requirement analysis, which will determine the business problems/questions and user information needs that will be solved using the data warehouse. Examples of business questions regarding a web site and that can be answered with a data warehouse are (Hu, & Cercone, 2004):

- Which parts of the web site attract more visitors?
- Which is the most visited page in a given period?
- How much time is spent by page in a session?
- What is the average length of a session (in pages visited and time spent)?
- What are the hours with more web traffic?

Besides these, we believe that a data warehouse can be very useful in answering more challenging questions related to the management of web sites:

- Which pages in a web site attract users and make them customers?
- How good are the recommendations provided by the web site to a user?
- How adequate are the content descriptors (e.g., keywords, categories, etc) of a web page?

According to (Kimball, & Ross, 2002), there are two techniques to obtain the requirements for a data warehouse: interviews and/or facilitated sessions. Interviews are easier to schedule and allow individual participation. Facilitated sessions are group sessions led by an objective facilitator and whose goal is the gathering of information. These sessions may reduce the time to collect the requirements, but they require more commitment from the participants.

Once we have the requirements for the data warehouse, we can model it. A data warehouse can be developed using three different models: Cube, Star and Snowflake (Kimball, & Ross, 2002). In the Cube model, the data are stored in a cube of information that is computationally represented by multidimensional arrays (Velasquez, & Palade, 2008). A data warehouse can also be implemented in a Relational Data Base Management System (RDBMS) by representing a cube through relational tables using two logical models/schemas: Star or Snowflake. The Star model consists of a fact table and a set of dimension tables. Each record/tuple in the fact table stores the facts for analysis and the pointers (foreign keys) to the dimension tables. The Snowflake model extends the Star model by normalizing the dimensions and representing explicitly them in multiple related tables.

Independently of the model, a data warehouse stores its data as facts and dimensions. Facts are the central entities for analysis and can consist of measurements or facts of a business process (e.g., the time spent on an access, the event that generated an access, etc). Dimensions provide contextual information for the facts and can be used to constrain and/or group data when performing data warehousing queries. The dimensions

can be organized hierarchically into levels. For example, a dimension date can be organized into month and/or year.

An example of data warehouse for web site is described in (Velasquez, & Palade, 2008). The data warehouse stores the usage data in a fact table and the content and structure data in one of the dimension tables. Implemented in a RDBMS, their data warehouse is mainly used to support offline and online recommendations in order to build adaptive web sites. Offline recommendations consist of hyperlinks to be added to or eliminated from the current site, and (key)words to be used as "words to write" in the current and future pages. Online recommendations consist in suggesting pages that can be of interest for each user. We also propose a data warehouse for a business intelligent system for an e-news web portal which is described later in our case study.

Extraction, Transformation, and Loading Process

The Extraction, Transformation and Loading (ETL) process groups techniques and tools to extract data from different sources, followed by their transformation in useful information and loading into a data warehouse (Kimball, & Caserta, 2004). The ETL process is very important in the development of a data warehouse because the data, which are loaded into a data warehouse, stem from different sources, types, formats, etc. In the following, we briefly describe the three stages of the process:

- **Extraction:** This stage involves extracting the data from their sources and storing them in the Data Staging Area (DSA) that is the local where the data are transformed in useful information before being loaded into the data warehouse. Usually, a directory in the file system or a relational database is used to implement the Data Staging Area;
- **Transformation:** It is considered the most complex and time consuming stage because it applies a series of pre-processing and transformation functions to the data in the DSA in order to obtain the useful information that will be loaded into the data warehouse. Pre-processing and transformation are performed based on the type of data previously defined to be loaded into the data warehouse;
- **Loading:** It loads the data into the data warehouse. Depending on the location of the DSA and data warehouse, this may be the simplest stage in the ETL process. If they are both on the same computer, we will just need to load the data warehouse with the data from the DSA. If the DSA is on an independent computer, we will have to define an interchange protocol to transmit data from the DSA to the data warehouse (Velasquez, & Palade, 2008).

These three stages can be combined with the techniques for pre-processing of web data, presented previously, in order to develop an ETL process for web data.

Online Analytical Processing

The information stored in a data warehouse is usually exploited by online analytical processing (OLAP). In this section, we briefly describe a few concepts about OLAP which are necessary for this chapter. A more complete overview is presented in (Malinowski, & Zimnyi, 2008).

In (Codd, Codd, & Salley, 1993), OLAP is defined as an approach to answer quickly multidimensional analytical queries. There are three types of OLAP implementations: MOLAP, ROLAP and HOLAP (Kimball, & Merz, 2000). We use Multidimensional OLAP (MOLAP) if the data cube is implemented in a MDBMS. On the other hand, if the data cube is implemented in a RDBMS, we use the Relational OLAP (ROLAP).

The Hybrid OLAP (HOLAP) is used if the data cube is implemented using both a MDBMS and a RDBMS. Some common operations used in OLAP analyses, independently of implementation, are presented below (Malinowski, & Zimnyi, 2008):

- **Slice:** Generating a sub-cube corresponding to a single value for one or more data fields of a dimension;
- **Dice:** The dice operation is a slice on two or more dimensions of a data cube;
- **Drill Down/Up:** Drilling down or up is an analytical technique to show the data among their levels of aggregation, going from the most summarized (up) to the most detailed (down);
- **Drill-Across:** Running queries involving more than one data cube;
- **Pivot:** Changing the dimensional orientation of a cube to show a particular face.

DEFINING PATTERN DISCOVERY AND ANALYSIS

Besides OLAP analyses, once the data are collected and stored in a data warehouse, we can also apply statistical, data mining and machine learning techniques on the data in order to discover useful patterns (Mitchell, 1997; Witten, & Frank, 2005). In web mining, the techniques most often used to extract patterns from web data are: statistical analysis, association rules, sequential patterns, clustering, classification and dependency modeling (Srivastava, Cooley, Deshpande, & Tan, 2000; Eirinaki, & Vazirgiannis, 2003). We briefly describe these techniques below:

- **Statistical Analysis:** We can use this technique to perform different types of descriptive statistical analyses (e.g., frequency, mean, median, etc) on variables such as pages, time spent on pages, length of sessions, and so forth;
- **Association Rules:** This technique can be used to find frequent patterns, associations and correlations among sets of items/pages;
- **Sequential Patterns:** This one is an extension of association rules in that it reveals patterns of co-occurrence incorporating the notion of time sequence;
- **Clustering:** We use this technique to group together a set of items or users with similar characteristics/behaviors;
- **Classification:** This one maps an item (e.g., a web page) into one of several predefined classes;
- **Dependency Modeling:** We can use this technique to build models which represent significant dependencies among the various variables in the web domain (e.g., web accesses).

After discovering patterns, we must conduct a further analysis in order to filter out the patterns that are not interesting. Then, the remaining patterns can be used by a wide range of applications, for example, a decision support system.

USING WEB MINING AS A PROCESS TO DEVELOP AN ADVANCED DECISION SUPPORT SYSTEM FOR AN E-NEWS WEB PORTAL

The goal of many web portals is to select, organize and distribute content (information, or other services and products) in order to satisfy its users/customers. The methods to support this process are to a large extent based on meta-data (such as keywords, categories, authors and other descriptors) that describe content and its properties. For instance, search engines often take into account keywords that are associated with the content to compute their relevance to a query. Likewise, the accessibility of content by navigation depends on their position in the structure of the portal,

which is usually defined by a specific meta-data descriptor (e.g., category). Meta-data is usually filled in by the authors who publish content in the portal. The publishing process, which goes from the insertion of content to its actual publication on the portal is regulated by a workflow. The complexity of this workflow varies: the author may be authorized to publish content directly; alternatively content may have to be analyzed by one or more editors, who authorize its publication or not. Editors may also suggest changes to the content and to the meta-data that describe it, or make those changes themselves.

In the case where there are many different authors or the publishing process is less strict, the meta-data may describe content in a way which is not suitable for the purpose of the portal, thus decreasing the quality of the services provided. For instance, a user may fail to find relevant content using the search engine if the set of keywords assigned to it are inappropriate. Thus, it is essential to monitor the quality of meta-data describing content to ensure that the collection of content is made available in a structured, inter-related and easily accessible way to the users.

In this section, we present a case study that demonstrates how the web mining process, described in the previous section, can be used to develop a web-based decision support system to monitor/manage the quality of meta-data describing content in an e-news web portal (Domingues, Soares, & Jorge, 2013). To develop the system, we follow the four steps of our web mining process: defining web data, pre-processing web data, web data warehousing, and defining pattern discovery and analysis.

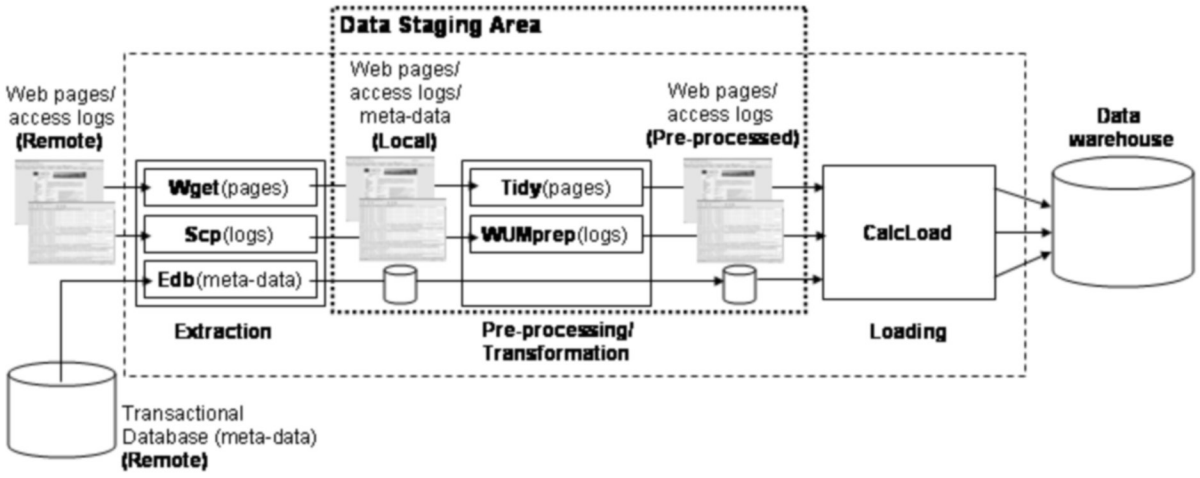
Defining and Pre-Processing Web Data

In this work, we use content, structure and usage data. As already stated, these are used as input for the most common web site management applications. To collect and pre-process these data, we have developed an ETL (Extraction, Transformation and Loading) tool.

The ETL tool is presented in Figure 3. One of the advantages of this tool is that it is developed as a composition of other different existing tools. Another advantage is that it runs as a batch process, without any manual intervention.

As the name indicates, the process is done in three steps: *extraction*, *pre-processing/transformation* and *loading*. In the *extraction* step, the process creates a local version of (the possibly remote) activity data. This local version is stored

Figure 3. The extraction, transformation and loading (ETL) tool



in the Data Staging Area (DSA), a simple directory in the file system. For this task, we use *Wget*[2], *Scp*[3] and *Edb*. ***Wget*** is a free software for retrieving remote files using HTTP, HTTPS and FTP, which are the most widely used Internet protocols. ***Scp*** is a software implementing the SCP protocol for secure copying of files between a local and a remote host or between two remote hosts. Finally, *Edb* is a SQL component developed by us to select meta-data from a transactional database and create a local version in text files.

In the following step, the local version of the activity data are pre-processed and transformed in useful information ready to compute the quality metrics and be loaded into a data warehouse. For web pages (contents), the process reads the HTML files, and writes clean and well-formed markup in XHTML format. For this task, we use *Tidy*[4]. This is an open source software and library for checking and generating clean and well-formed XML/XHTML/HTML files. The pre-processing of the access logs consists of merging the log files, removing irrelevant requests and/or data fields, removing robot requests, and identifying users and sessions for the local version of the access logs. We use *WUMPrep*[5], a collection of Perl programs supporting data preparation for data mining of web logs. Regarding the meta-data, we do not pre-process and transform them given that our goal is to evaluate their quality.

At this point, we are ready to compute the metrics and load them, together with the activity data, into the data warehouse. Therefore, for the *loading* step, we have implemented a R-based component, called *CalcLoad*, that computes the quality metrics and load them into the data warehouse. R[6] is an integrated software suite for data manipulation, calculation and graphical display. The data warehouse used to store the data and the metrics are presented in the next sections.

Web Data Warehousing

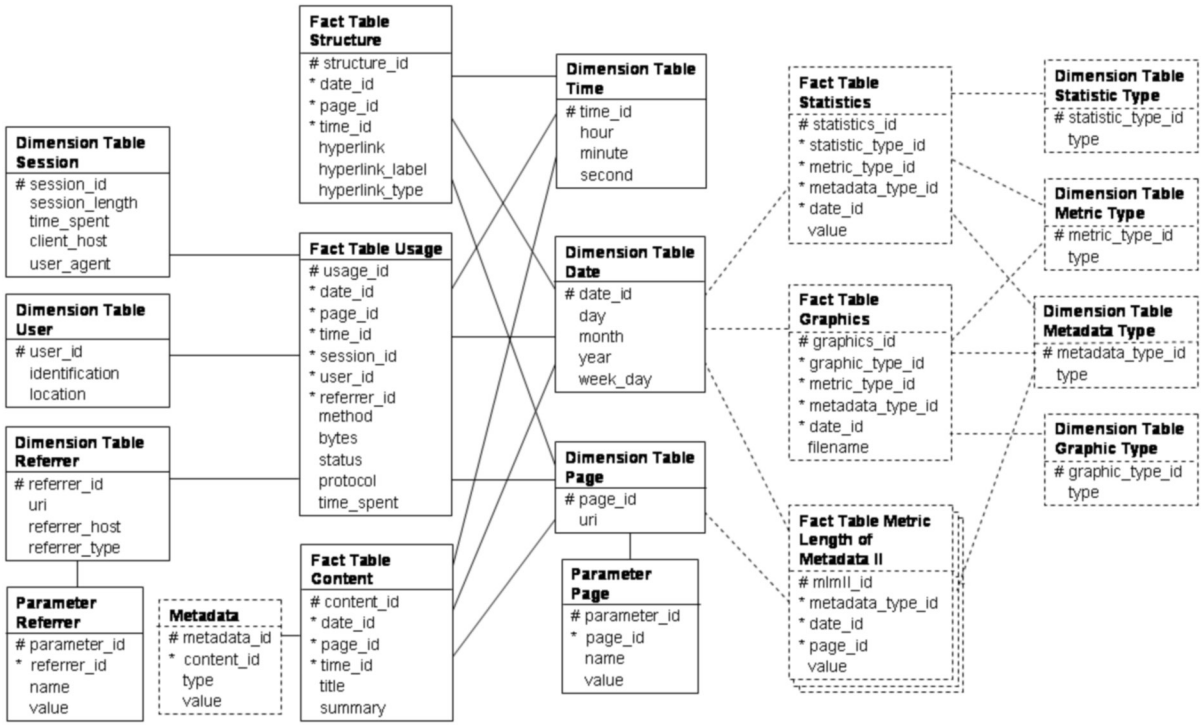
In order to provide a suitable repository for data and metrics, we have extended the data warehouse proposed in (Domingues, Jorge, Soares, Leal, & Machado, 2007) by including tables which make possible the storage of meta-data, metrics and macro indicators (i.e., statistics and graphics).

The extended star schema for the data warehouse is presented in Figure 4. The original version of the data warehouse only contains fact tables to store data related to structure, usage and content of a web site. As we want to provide analyses which are based on meta-data, metrics and macro indicators (see next section), we need to have tables to store these data. In this new version, we have extended the star schema by including such tables. The addition is represented by the dashed tables in Figure 4.

In the fact table *Structure*, we store each hyperlink in the web site, and consequently, the hierarchy organization of the site. The information from the web logs are stored in the fact table *Usage*. Finally, we store the content (i.e., title and summary) and its meta-data (i.e., type and value) in the tables *Content* and *Metadata*. In this new version of the data warehouse, the addition of the table *Metadata* and its relationship with the table *Content* allow the storage of different meta-data which belong to a specific content item.

In the star schema, each metric has its own fact table to store its value and information related to it (e.g., type of meta-data assessed by the metric, page which the meta-data are associated to, etc). In Figure 4, we can see an example of fact table for the metric *Length of meta-data 2*, which consists in computing the number of words in a meta-data field (see next section). It stores the type of meta-data that is assessed (foreign key *metadata_type_id*), when the metric is calculated

Figure 4. Star schema of the data warehouse emphasizing the tables used in the computation and storage of the metric Length of meta-data 2



(foreign key *date_id*), the web page which the metric is associated to (foreign key *page_id*) and the *value* of the metric.

The statistical indicators and graphics are stored in the fact tables *Statistics* and *Graphics*, which are very close each other in terms of structure (Figure 4). The fact table *Statistics* stores the type of statistical indicator (foreign key *statistic_type_id*) and the *value* for the statistic. The fact table *Graphics* stores the type of graphical representation (foreign key *graphic_type_id*) and the *file name* for the graphic. Additionally, both tables also store the metric used by the statistics or graphics (foreign key *metric_type_id*), the type of meta-data assessed by the metric (foreign key *metadata_type_id*) and the date of computation (foreign key *date_id*). The types of statistical indicators, metrics, meta-data and graphics are stored, respectively, in the dimension tables *Statistic Type, Metric Type, Metadata Type and Graphic Type.*

Defining Pattern Discovery and Analysis

In the step of pattern discovery and analysis, we have focused on statistical analysis and association rules to compute the metrics for monitoring the quality of meta-data. We have proposed 31 metrics for measuring the quality of content meta-data, which were designed based on data quality principles (Spiliopoulou, & Pohle, 2001; Pipino, Lee, & Wang, 2002; Moorsel, 2001). Table 2 presents a few examples for illustration purposes. The complete list of metrics for measuring the quality of content meta-data is presented in the Appendix.

The functions used to compute the metrics can be based on very simple statistics or more complex methods. For instance, the metric *Length of meta-data 2* is computed simply by counting the number of words in a meta-data field. Metrics

Table 2. Name and description of a few metrics

Name: *Length of meta-data 2* **Description:** Number of words in a meta-data field. Extremely large or small values may indicate an inadequate choice of meta-data to represent the content.
Name: *Association between meta-data values* **Description:** The confidence level of an association rule *X* -> *Y* is an indicator of whether the set of values *X* makes the set of values *Y* redundant or not. The higher the value, the more redundant *Y* is expected to be. This may indicate that implicit practices in the description of content have been developed.
Name: *Frequency in search* **Description:** Number of meta-data values in the web access logs. For instance, the frequency of a search using a given keyword. If such a keyword is searched for often, probably it will have a high interpretability.
Name: *Redundancy of meta-data values* **Description:** Conditional probability *P(x|y)*, where *x* is one meta-data value of a content, and *y* is another one. High values may mean that y makes *x* redundant. This may indicate that implicit practices in the description of content have been developed.

based on simple frequencies, such as the *Frequency in search* (Table 2), are quite common. Alternatively, metrics can be based on probabilities. The *Redundancy of meta-data values* metric is based on the conditional probability of having a value *x*, in the description of content, given that another value *y* is used (Table 2). An example of a more complex method is given by association rules (Agrawal, & Srikant, 1994), which are used to compute the *Association between meta-data values* metric (Table 2). The computation of the metrics is usually based on the meta-data. However, in some cases the information about usage can also be used, such as in the case of the *Frequency in search* metric.

Here, we illustrate the computation of the metric *Length of meta-data 2*. This metric uses data that are in the fields *type* and *value* from the table *Metadata*, *uri* from the table *Page*, and *day*, *month* and *year* from the table *Date* that additionally use the table *Content* to establish a relationship among them. The data in the fields *day*, *month* and *year* are used to indicate which version of the page and its meta-data must be retrieved. It is necessary because the data warehouse stores periodically the content of the web site to make possible the analysis of its evolution.

The metric is stored in the fact table *Metric Length of Metadata 2*. The table stores the type of meta-data that is assessed (foreign key *meta-data_type_id*), when the metric is calculated (foreign *key_date_id*), the web page which the meta-data are associated to (foreign *key_page_id*) and the *value* of the metric.

Once we have the metric *Length of meta-data 2* calculated, we can compute its statistical indicators and graphics. First, we retrieve all values from the fact table *Metric Length of Metadata 2*. Then, we use the retrieved values to compute the statistical indicators (for this metric, minimum and maximum value) and plot graphics showing the evolution in time of the values. The statistics and graphics are stored in the fact tables *Statistics* and *Graphics.*

EVALUATING OUR DECISION SUPPORT SYSTEM IN AN E-NEWS WEB PORTAL

The developed decision support system was deployed in the PortalExecutivo (PE), a Portuguese e-news web portal targeted to business executives. The business model of the portal is subscription-based, which means that only paying users have full access to content through web login. However, some content is freely available and users can freely browse the structure of the site. Content is provided not only by PE but also by a large number of partners. The goal of PE is to facilitate the access of its customers to relevant content. Value is added to the contributed content by structuring and interrelating them. This is achieved by filling in a rich set of meta-data fields, including keywords, categories, relevant companies, authors, among others. Thus, monitoring the meta-data fields, detecting unusual values and correcting them, is very important to PE because it can guarantee the

quality of the meta-data and, consequently, add value to the content.

An example of a particularly important meta-data is keyword, which characterize the content of a web page or site, and are used by users in their search process. Since the access to a content (e.g., using a search engine) is affected by the quality of the keywords describing the content, the developed system was applied to monitor the quality of this meta-data in the PE. The keywords monitored are relative to the period April/September 2004[7]. In this period we have 17,196 content items and 124,287 web accesses recorded.

Figure 5 presents a very simple example that illustrates the type of analysis that can be carried out with the system. The metric represented in the figure is the number of keywords which are used only once. Higher values of keywords with frequency equal 1 may indicate that the potential of the keywords to interrelate content from different sources is not being adequately exploited or that these keywords with frequency equal 1 are typographical errors.

The results obtained with the system are not only useful to detect data quality problems but also to trigger corrective actions and monitor them. Figure 6 shows that in April more than 50% of content did not have any keyword filled in. This reduces the probability that these contents will be returned by the search engine of the web portal. To address this problem, the PE decided to implement a semi-automatic procedure to support the process of filling in keywords. The same figure shows that this caused a steady reduction in the number of contents without keywords, thus improving the quality of the meta-data.

The two described metrics are quite simple. More complex metrics may be interesting and can be implemented. For instance, the Association between meta-data values metric uses the confidence of association rules to determine keywords more frequently used together. The developed system collects from the data warehouse the keywords of each content as baskets of items. Then, it runs an association rules algorithm on the baskets to generate the associations among the keywords.

Figure 5. Evolution of the number of keywords with frequency 1 (Metric: Singleton meta-data values)

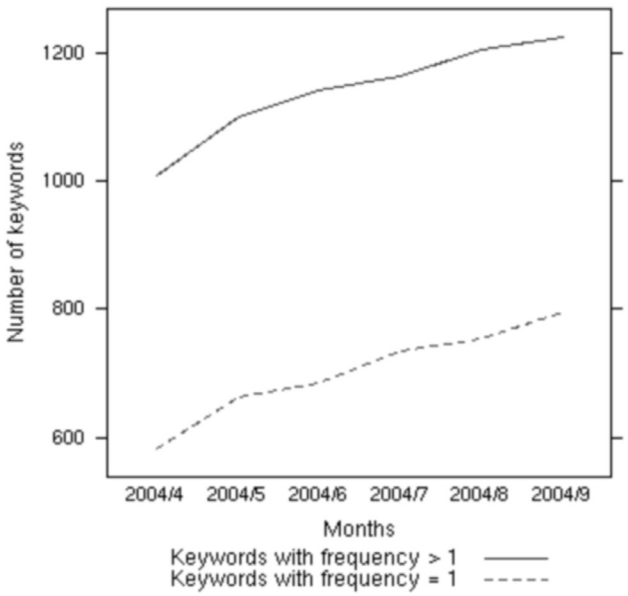
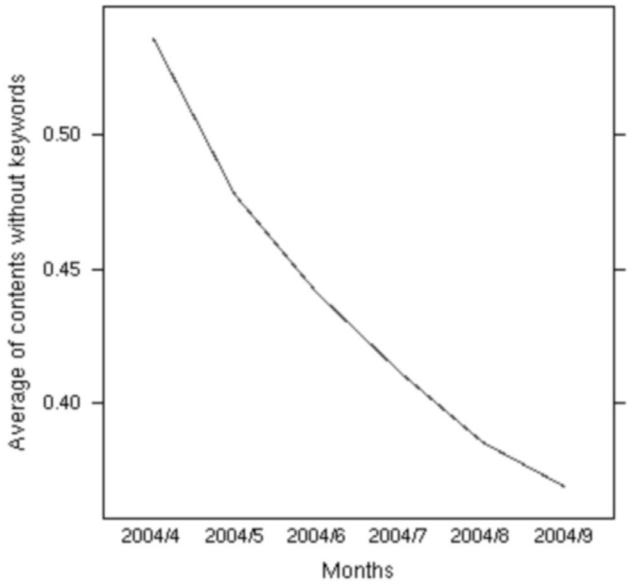


Figure 6. Evolution of the number of keywords not filled in (Metric: Empty meta-data field)



The current version of the system only provides a graphical representation of the associations among the keywords. In Figure 7, we see such a representation showing that often a general keyword (e.g., fiscality - fiscalidade) is associated with a more specific one (e.g., international taxation – tributação internacional). This implicit structure of the keywords, unveiled by the discovered association rules, enables the detection of incorrect descriptions.

Figure 7. Relationship between keywords obtained using association rules (Metric: Association between meta-data values)

FUTURE RESEARCH DIRECTIONS

As future work, we plan to use our proposed web mining process to develop other advanced decision support system for web sites. Regarding our system, developed for monitoring/managing the quality of meta-data, we plan to apply other statistical and data mining techniques to improve the quality assessment process. For instance, clustering methods (Velasquez, & Palade, 2008) can be used to obtain groups of authors with similar behaviors in terms of meta-data quality. This not only enables a different perspective on their publishing process but also different corrective actions can then be taken upon different groups.

CONCLUSION

In this chapter, we proposed to use web mining as a process to develop advanced decision support systems in order to support the monitoring/management activities of web sites. We described the web mining process as a sequence of four steps for the development of advanced decision support systems: defining web data, pre-processing web data, web data warehousing, and defining pattern discovery and analysis. By following this sequence, we developed an advanced decision support systems to monitor/manage the quality of meta-data in an e-news web site. With such a development, we concluded that by using our proposed process, we can integrate data mining and business intelligence techniques in order to develop more advanced decision support systems.

ACKNOWLEDGMENT

This work was supported by the grants 2011/19850-9 and 2012/13830-9, Sao Paulo Research Foundation (FAPESP).

REFERENCES

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of Twentieth International Conference on Very Large Data Bases* (pp. 487-499). Academic Press.

Anand, S. S., & Mobasher, B. (2003). Intelligent techniques for web personalization. In Intelligent Techniques for Web Personalization (LNCS), (vol. 3169, pp. 1-36). Berlin: Springer.

Azevedo, A., & Santos, M. F. (2011). A Perspective on Data Mining Integration with Business Intelligence. In A. Kumar (Ed.), *Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains* (pp. 109–129). Hershey, PA: IGI Publishing.

Berendt, B., Mobasher, B., Nakagawa, M., & Spiliopoulou, M. (2002). The impact of site structure and user environment on session reconstruction in web usage analysis. In Proceedings of WEBKDD 2002 - Mining Web Data for Discovering Usage Patterns and Profiles (pp. 159-179). Springer.

Berners-Lee, T., & Connolly, D. (1995). *Hypertext markup language - 2.0.* Technical report rfc 1866. Retrieved July 12, 2013, from http://www.ietf.org/rfc/rfc1866.txt

Berners-Lee, T., Fielding, R., & Masinter, L. (1998). *Uniform resource identifiers (uri): generic syntax.* Technical report rfc 2396. Retrieved July 12, 2013, from http://www.ietf.org/rfc/rfc2396.txt

Bonchi, F., Giannotti, F., Gozzi, C., Manco, G., Nanni, M., & Pedreschi, D. et al. (2001). Web log data warehousing and mining for intelligent web caching. *Data & Knowledge Engineering*, *39*(2), 165–189. doi:10.1016/S0169-023X(01)00038-6

Chakrabarti, S. (2000). Data mining for hypertext: A tutorial survey. *ACM SIGKDD Explorations Newsletter*, *1*(2), 1–11. doi:10.1145/846183.846187

Chakrabarti, S., Dom, B. E., Kumar, S. R., Raghavan, P., Rajagopalan, S., & Tomkins, A. et al. (1999). Mining the link structure of the world wide web. *IEEE Computer*, *2*(8), 60–67. doi:10.1109/2.781636

Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and olap technology. *SIGMOD Record*, *26*(1), 65–74. doi:10.1145/248603.248616

Chen, M.-S., Park, J. S., & Yu, P. (1996). Data mining for path traversal patterns in a web environment. In *Proceedings of the Sixteenth International Conference on Distributed Computing Systems* (pp. 385). Washington, DC: IEEE Computer Society.

Claypool, M., Brown, D., Le, P., & Waseda, M. (2001). Inferring user interest. *IEEE Internet Computing*, *5*(6), 32–39. doi:10.1109/4236.968829

Codd, E. F., Codd, S. B., & Salley, C. T. (1993). *Providing olap (on-line analytical processing) to user-analysis: An it mandate* (White Paper). Academic Press.

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, *1*(1), 5–32. doi:10.1007/BF03325089

Domingues, M. A., Jorge, A. M., Soares, C., Leal, J. P., & Machado, P. A. (2007). A Data Warehouse for Web Intelligence. In *Proceedings of EPIA 2007 - 13th Portuguese Conference on Artificial Intelligence* (pp. 487-499). Springer.

Domingues, M. A., Soares, C., & Jorge, A. M. (2013). Using statistics, visualization and data mining for monitoring the quality of meta-data in web portals. *Information Systems and e-Business Management, 11*(4), 569-595.

Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for web personalization. *ACM Transactions on Internet Technology*, *3*(1), 1–27. doi:10.1145/643477.643478

Etzioni, O. (1996). The world-wide web: Quagmire or gold mine? *Communications of the ACM*, *39*(11), 65–68. doi:10.1145/240455.240473

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining* (pp. 1–34). Menlo Park, CA: American Association for Artificial Intelligence.

Hu, X., & Cercone, N. (2004). A data warehouse/online analytic processing framework for web usage mining and business intelligence reporting. *International Journal of Intelligent Systems*, *19*(7), 585–606. doi:10.1002/int.20012

Inmon, W. H. (2005). *Building the Data Warehouse*. New York, NY: John Wiley & Sons, Inc.

Joshi, K. P., Joshi, A., & Yesha, Y. (2003). On using a warehouse to analyze web logs. *Distributed and Parallel Databases*, *13*(2), 161–180. doi:10.1023/A:1021515408295

Kimball, R., & Caserta, J. (2004). *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. New York, NY: John Wiley & Sons, Inc.

Kimball, R., & Merz, R. (2000). *The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse*. New York, NY: John Wiley & Sons, Inc.

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. New York, NY: John Wiley & Sons, Inc.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations: Newsletter of the Special Interest Group on Knowledge Discovery and Data Mining*, *2*(1), 1–15. doi:10.1145/360402.360406

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling the web for emerging cyber-communities. *Computer Networks*, *31*(11-16), 1481–1493. doi:10.1016/S1389-1286(99)00040-7

Malinowski, E., & Zimnyi, E. (2008). *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications (Data-Centric Systems and Applications)*. Springer Publishing Company.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Moorsel, A. V. (2001). Metrics for the internet age: quality of experience and quality of business. In *Proceedings of Fifth Performability Workshop*. HP Laboratories.

Pabarskaite, Z., & Raudys, A. (2007). A process of knowledge discovery from web log data: Systematization and critical review. *Journal of Intelligent Information Systems*, *28*(1), 79–104. doi:10.1007/s10844-006-0004-1

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: bringing order to the web.* Technical report Stanford InfoLab. Retrieved November 20, 2013, from http://ilpubs.stanford.edu:8090/422/

Peterson, E. T. (2004). *Web Analytics Demystified: A Marketer's Guide to Understanding How Your Web Site Affects Your Business*. Celilo Group Media.

Pipino, Y. W., Lee, L. L., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*,*45*(4),211–218.doi:10.1145/505248.506010

Pirolli, P., Pitkow, J., & Rao, R. (1996). Silk from a sow's ear: Extracting usable structures from the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 118-125). New York, NY: ACM.

Raju, G. T., & Satyanarayana, P. S. (2008). Knowledge discovery from web usage data: A complete preprocessing methodology. *International Journal of Computer Science and Network Security, 8*(1).

Spiliopoulou, M., & Pohle, C. (2001). Data mining for measuring and improving the success of web sites. *Data Mining and Knowledge Discovery*, *5*(1-2), 85–114. doi:10.1023/A:1009800113571

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, *1*(2), 12–23. doi:10.1145/846183.846188

Tanasa, D., & Trousse, B. (2004). Advanced data preprocessing for intersites web usage mining. *IEEE Intelligent Systems*, *19*(2), 59–65. doi:10.1109/MIS.2004.1274912

Velasquez, J. D., & Palade, V. (2008). Adaptive Web Sites: A Knowledge Extraction from Web Data Approach (Vol. 170). Amsterdam, Netherlands: IOS Press.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). San Francisco, CA: Morgan Kaufmann Publishers Inc.

Wookey, L., & Geller, J. (2004). Semantic hierarchical abstraction of web site structures for web searchers. *Journal of Research and Practice in Information Technology*, *6*(1), 23–34.

Zhang, Y., Zincir-Heywood, N., & Milios, E. (2004). World wide web site summarization. *Web Intelligence and Agent Systems*, *2*(1), 39–53.

ADDITIONAL READING

Burstein, F., & Holsapple, C. W. (2008). *Handbook on Decision Support Systems*. Springer-Verlag Berlin.

Das, R., & Turkoglu, I. (2009). Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. *Expert Systems with Applications*, *36*(3), 6635–6644. doi:10.1016/j.eswa.2008.08.067

Jao, C. S. (2010). *Decision Support Systems*, Publisher: InTech.

Kim, J. H., Sohn, J. G., Yun, M., Oh, G. Y., Choi, H. I., & Choi, H. A. (2013). Design of a web-based decision support system for service portfolios in heterogeneous radio access network environments. *Journal of Network and Systems Management*, *21*(3), 353–383. doi:10.1007/s10922-012-9239-z

Lappas, G. (2009). Machine learning and web learning: Methods and applications in societal benefit areas. In *Data Mining Applications for Empowering Knowledge Societies* (pp. 76–95). Hershey, PA: IGI Publishing.

Liu, B. (2007). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer-Verlag Berlin Heidelberg.

Markov, Z., & Larose, D. T. (2007). *Data Mining in the Web: Uncovering Patterns in Web Content, Structure, and Usage*. Koboken, New Jersey: Wiley-Interscience. doi:10.1002/0470108096

Ponis, S. T., & Christou, I. T. (2013). Competitive intelligence for SMEs: A web-based decision support system. *International Journal of Business Information Systems*, *12*(3), 243–258. doi:10.1504/IJBIS.2013.052449

Power, D. J., & Kaparthi, S. (2002). Building web-based decision support systems. *Studies in Informatics and Control*, *11*, 291–302.

Rupnik, R., & Kukar, M. (2007). Data mining based decision support system to support association rules. *Elektrotehniski vestnik 74(4)*, 195-200.

Tsolis, D., Paschali, K., Tsakona, A., Ioannou, Z. M., Likothanasis, S., & Tsakalidis, A. et al. (2013). Development of a clinical decision support system using AI, medical data mining and web applications. In *Proceedings of the 14th International Conference on Engineering Applications of Neural Networks* (pp 174-184), Halkidiki, Greece. doi:10.1007/978-3-642-41016-1_19

Wright, A., & Sittig, D. (2008). A framework and model for evaluating clinical decision support architectures. *Journal of Biomedical Informatics*, *41*(6), 982–990. doi:10.1016/j.jbi.2008.03.009 PMID:18462999

KEY TERMS AND DEFINITIONS

Association Rules: A technique to find frequent patterns, associations and correlations among sets of items.

Business Intelligence (BI): A set of technologies to collect and store data, analyze it using analytical tools, and deliver information and/or knowledge, facilitating reporting and querying.

Data Mining: Application of specific algorithms for extracting patterns from data.

Data Warehouse: A subject oriented, integrated, time-variant, and non-volatile collection of data in support of management decision making process.

Decision Support System: A computer-based information system that supports business and organizational decision-making activities.

Extraction, Transformation, and Loading (ETL) Process: Techniques to extract data from different sources, followed by their transformation in useful information and loading into a data warehouse.

OLAP: An approach to answer quickly multidimensional analytical queries.

Web Mining: Application of data mining techniques to automatically discover and extract information from web sites.

ENDNOTES

1 http://www.adobe.com/products/flash/
2 http://www.gnu.org/software/wget/
3 http://www.openssh.org/
4 http://tidy.sourceforge.net/
5 http://hypknowsys.sourceforge.net/wiki/Web-Log_Preparation_with_WUMprep/
6 http://www.r-project.org/
7 We only have permission to publish results for this period of time

APPENDIX

Table 3. Name and description of metrics for measuring the quality of content meta-data

Name: *Number of shadow contents* **Description:** A content c1 is shadow of a content c2 if the set of meta-data values in c2 is a super-set of c1. The access to a content that is shadow of other contents may be very difficult using the meta-data.
Name: *Association between meta-data values* **Description:** The confidence level of an association rule *X -> Y* is an indicator of whether the set of values *X* makes the set of values *Y* redundant or not. The higher the value, the more redundant *Y* is expected to be. This may indicate that implicit practices in the description of content have been developed.
Name: *Frequency in search* **Description:** Number of meta-data values in the web access logs. For instance, the frequency of a search using a given keyword. If such a keyword is searched for often, probably it will have a high interpretability.
Name: *Redundancy of meta-data values* **Description:** Conditional probability *P(x|y)*, where *x* is one meta-data value of a content, and *y* is another one. High values may mean that y makes *x* redundant. This may indicate that implicit practices in the description of content have been developed.
Name: *Length of meta-data 1* **Description:** Number of characters in a meta-data field. Extremely large or small values may indicate an inadequate choice of meta-data to represent the content.
Name: *Length of meta-data 2* **Description:** Number of words in a meta-data field. Extremely large or small values may indicate an inadequate choice of meta-data to represent the content.
Name: *Length of title/summary 1* **Description:** Number of characters in the title/summary. Large or small values mean that the choice of the title or summary may not be adequate.
Name: *Length of title/summary 2* **Description:** Number of words in the title/summary. Large or small values mean that the choice of the title or summary may not be adequate.
Name: *Length of title/summary 3* **Description:** Number of phrases in the title/summary. Large or small values mean that the choice of the title or summary may not be adequate.
Name: *Extreme frequency of meta-data values* **Description:** Number of contents which contain a given meta-data value. High values may indicate that the number of contents selected by using such a meta-data will be very high.
Name: *Extreme frequency of meta-data values, by editor/author* **Description:** Number of contents, grouped by editor/author, which contain a given metadata value.
Name: *Shared meta-data values* **Description:** Number of meta-data values which are used by at least two different contents. The shared values allow the relationship among contents.
Name: *Degree of sharing among editors/authors* **Description:** Number of different editors/authors who use/share a same meta-data value.
Name: *Empty meta-data field* **Description:** Number of contents with a given meta-data field not filled in. If a meta-data is used to find a content but the meta-data field is not filled in, the content will not be found.
Name: *Empty meta-data field, by editor/author* **Description:** Number of contents, grouped by editors/authors, with a given meta-data field not filled in.
Name: All empty meta-data fields **Description:** Number of contents with all meta-data fields not filled in. Contents without any meta-data may indicate errors of publication.

continued on following page

Table 3. Continued

Name: *All empty meta-data fields, by editor/author* **Description:** Number of contents, grouped by editors/authors, with all meta-data fields not filled in. Here, we can evaluate the not filling in of meta-data by editors/authors.
Name: *Length of meta-data values* **Description:** Number of characters in the value of a meta-data. Extremely large or small values may indicate an inadequate choice of meta-data values to represent the content.
Name: *Quantity of meta-data values* **Description:** Number of different values which are used as meta-data.
Name: *Quantity of meta-data values per content* **Description:** Number of different values which are used as meta-data per content. Lower quantities may indicate regular procedures of filling in. Higher quantities may indicate a careful filling in, but maybe with the insertion of values with low description.
Name: *Singleton meta-data values* **Description:** Meta-data values which are used only once. This metric may indicate typographical errors.
Name: *Repeated meta-data values* **Description:** Number of repeated meta-data values in a content.
Name: *Contents with repetition* **Description:** Number of contents with the same meta-data values.
Name: *Existence in another field* **Description:** Meta-data values which appear in another data field. For instance, a metadata value x may be a good choice if it also appears in the title/summary of a content.
Name: *Isolated usage* **Description:** Number of editors/authors who always use the same meta-data values.
Name: *Latency of free access* **Description:** If the free access date for a content is set to a date that comes quite later the publication date (for instance, more than 1 year), the value for this meta-data may not be adequate.
Name: *Invalid free access range* **Description:** The free access range is invalid when the start date for free access is set to a date that comes after the end date for free access.
Name: *Shorted free access range* **Description:** This metric indicates whether there is a small difference between the start and the end date for the free access or not.
Name: *Extreme price* **Description:** Higher or lower selling prices for a content may indicate an inadequate choice for this meta-data.
Name: *Invalid price* **Description:** The selling price of a content is invalid when it is negative.
Name: *Depth of the hierarchy* **Description:** This metric shows the depth of the hierarchy of a web site. A great number of levels in the hierarchy means that will be difficult to find a content if it is in the lowest levels of this hierarchy.