# MANAGING WEB-BASED INFORMATION

Marco Scotto, Tullio Vernazza
*DIST – Università di Genova,Via Opera Pia, 13, I-16145 Genova, Italy*

Alberto Sillitti, Giancarlo Succi
*Libera Università di Bolzano, Piazza Domenicani, 3, I-39100 Bolzano, Italy*

Keywords: Web Mining, Information Retrieval

Abstract: The heterogeneity and the lack of structure of World Wide Web make automated discovery, organization, and management of Web-based information a non-trivial task. Traditional search and indexing tools provide some comfort to users, but they generally provide neither structured information nor categorize, filter, or interpret documents in an automated way. In recent years, these factors have prompted the need for developing data mining techniques applied to the web, giving rise to the term "Web Mining". This paper introduces the problem of web data extraction and gives a brief analysis of the various techniques to address it. Then, News Miner, a tool for Web Content Mining applied to the news retrieval is presented.

## 1 INTRODUCTION

The World Wide Web has become a huge source of information, but its content cannot be manipulated in a general way due to two main issues:

- Finding relevant information is a difficult task because the web is unstructured. Search engines, such as Altavista, Google, Lycos and many others, provide some comfort to users, but their query facilities are often limited and the results come as HTML pages
- Most of the information, present on the web, is stored as HTML pages. HTML is a semi-structured format designed to describe the layout of web pages not their content, and machines hardly process HTML.

These factors have prompted the need for developing data mining techniques applied to the web, giving rise to the term "Web Mining". This paper introduces the problem of the extraction of information from the web and analyzes the various techniques to approach it. Then, it presents News Miner, a tool for news extraction, integration and presentation, based on Web Content Mining. News Miner is a server side application that periodically scans a set of news sites, integrates news building a repository, and makes them available to an application server that use such information to build user-specific web pages. In particular, it organizes HTML documents, which are semi-structured, into structured XML documents, using XSLT (eXtensible Stylesheet Language Transformation) and XPath (XML Path Language). This paper is organized as follows: section 2 introduces Web Mining; section 3 analyzes the different types of Web Mining; section 4 describes the design and the implementation of News Miner; finally, section 5 draws the conclusions.

## 2 WEB MINING

Web Mining can be broadly defined as "the discovery and analysis of useful information from the World Wide Web" (Madria *et al.*, 1999). In Web Mining, data can be collected at different levels: server side, client side, proxy servers, or obtained from an organization's database. For instance, data can be stored in browser caches or in cookies at client level, and in access log files at server or proxy level. Web Mining can be decomposed into four



Scotto M., Vernazza T., Sillitti A. and Succi G. (2004).
MANAGING WEB-BASED INFORMATION.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 575-578
DOI: 10.5220/0002633005750578

subtasks, according to Etzioni (Etzioni, 1996). Each task is discussed in the next paragraphs.

## 2.1 Information Retrieval

Information Retrieval deals with automatic discovery of all relevant documents satisfying a specific query. Most of the work on information retrieval focuses on automatic indexing of web documents. However, web pages indexing is not a trivial task if compared to database indexing where there are well-defined techniques. The huge number of web pages, their heterogeneity, and frequent changes in number and content make this task very difficult. At present there are several search engines for querying and retrieve web documents, each one has a unique interface and a database, which covers a different fraction of the Web. Their indexes have been created and constantly updated by web robots, which scan millions of web pages and store an index of the words in the documents.

## 2.2 Information Extraction

Once the documents have been retrieved, the challenge is the automated extraction of knowledge from the source without any human effort. At present most of the work in information extraction is carried out by wrappers built around web sources. A wrapper is a special program, which accepts queries about information present in the pages of the source, extracts the requested information and returns the result. But it is impractical to build wrappers for web sources by hand for several reasons: the number of web pages is very large, a lot of new pages are frequently added and the format of web pages often changes. Ashish and Knoblock (Ashish and Knoblock, 1997) propose an approach to semi-automatically generation of wrappers for Web sources.

## 2.3 Generalization

Once automated the discovery and extraction processes from web pages, the next step is the generalization from the experience. This phase involves pattern recognition and machine learning techniques. The bigger obstacle in learning about web is the large amount of unlabelled data. Many data mining techniques require inputs labelled as positive or negative examples with respect to some concept. Fortunately, clustering techniques do not require labelled inputs and have been applied successfully to large collections of documents. Other techniques, used in this phase, are association rules. They allow the discovery of all associations and correlations among data items where the presence of one set of items in a transaction implies, with a certain degree of confidence, the presence of other items.

## 2.3 Analysis

Analysis is a data-driven problem where humans play an important role for validation and interpretation of the results. Once patterns have been discovered, analysts need suitable tools to understand, visualize, and interpret these patterns. One technique is represented by OLAP (On Line Analytical Process), which uses data cube structure for simplifying visualization of multidimensional data. Some others (Mobasher *et al.*, 1997) proposed an SQL-like language for querying the discovered knowledge.

# 3 WEB MINING CATEGORIES

Web Mining includes three areas, based on which part of the Web mine:

- Web Content Mining (WCM),
- Web Structure Mining (WSM),
- Web Usage Mining (WUM).

The distinctions among the above categories are not clear-cut; the three Web Mining tasks could be used in isolation or combined in an application. An overview of each category follows.

## 3.1 Web Content Mining

The aim of the WCM is the automation of the process of information discovery and extraction from Web documents and services. Mainly, there are two approaches to solve this problem (Cooley *et al.*, 1997):

1. *Agent Based approach*: “it involves artificial intelligence systems that can act autonomously or semi-autonomously on behalf of a particular user, to discover and organize Web-Based information”.
2. *Database approach*: “it organizes heterogeneous and unstructured or semi-structured data into more structured data, such as relational database, and using standard database querying mechanism and

data mining techniques to access and analyze this information".

## 3.2 Web Structure Mining

Web Structure Mining (Kleinberg, 1998) tries to extract information from the link structures of the web. The web can be modelled as a graph where Web pages are the vertices and hyperlinks are the edges. WSM categorizes Web pages and extract information, such as the similarity and relationship among different Web sites. For example, links pointing to a document offer an index of the popularity of the document (*authority site*), while links coming out of a document offer an index of the richness or the variety of topics covered in a document (*hub site*).

## 3.3 Web Usage Mining

While the two others techniques use the real or primary data on the web, Web Usage Mining (Srivastava *et al.*, 2000) mines secondary data generated by users interactions with the web. Web usage data includes data regarding web server access, logs, proxy server logs, browser logs, and any other data generated by the interaction of the users and the web. WUM is the "process of applying data mining techniques to the discovery of usage patterns from Web data". The knowledge extracted can be used to achieve different goals such as service personalization, site structure, and web server performance improvement.

# 4 NEWS MINER

News Miner is a server side application that periodically scans a set of news sites defined by the system administrator, integrates news building a repository, and makes them available to an application server. News Miner is based on the Service Oriented Architecture (Sillitti *et al.*, 2002); in particular, it exploits news services available on Internet to provide an integrated news service. The system is completely written in Java, using open source libraries (Apache Xalan, Apache Xerces, etc.), and standard W3C communication protocols (HTML, XML, etc.). The architecture includes three main components (Fig. 2): a *News Retriever*, a *News Repository*, and a *Data Provider*.

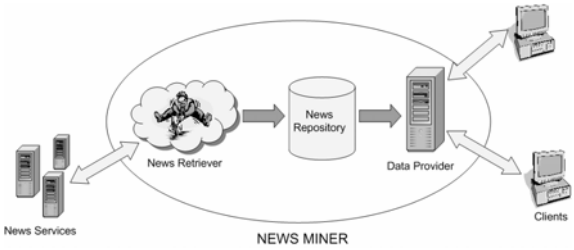


Figure 1: News Miner architecture

## 4.1 News Retriever

The *News Retriever* reads the configuration file and queries the web sites of selected news providers. It extracts news information, classifies news and stores them into a *News Repository*. This phase includes two main subtasks: *News Extraction* and *News Integration*.

The *News Extractor* downloads news web pages, extracts data and produces an XML document. The downloaded documents are often not well-formed HTML documents. Errors are corrected using an HTML lint that produce XHTML (eXtensible HTML) documents which are always well formed. An XHTML document is an XML document that can be processed using standard processing tools such as XSLT processors. Then, the *News Extractor* applies an XSLT transformation to the document to extract relevant data (Fig. 3) that the *News Integrator* collects as XML documents. The process is performed for each news provider that supplies the selected category, according to the configuration file. System administrator provides the transformation style sheets for each news provider at configuration time. The *News Integrator* performs categories merging of different news providers. Technologies involved in this step are DOM (Document Object Model) and XPath. At the end of the two processes, the system produces a news repository: a set of XML documents analyzable through standard data mining techniques.

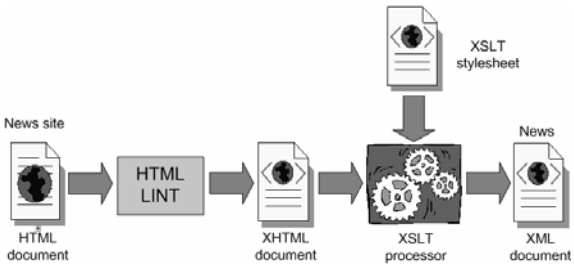


Figure 2: News Extractor overview

## 4.2 News Repository

The *News Repository* stores all the retrieved news and provides them to the *Data Provider*. News is stored as a set of XML documents. XML format is chosen because it is the standard interchange format for representing structured information.

## 4.3 Data Provider

The *Data Provider* (Fig. 4) sends data to different kind of clients (e.g., web browsers, handhelds, smart phones, etc.) adapting the content to the specific features of the client. The *Data Provider* is an *Application Server* that retrieves data from the *News Repository* and creates a presentation layout for sending to clients. The *Data Provider* adapts the format of the content to the specific client to produce the best results as possible. As instance, pages in the HTML format are provided to web browsers; the same set of pages is provided to WAP (Wireless Application Protocol) enabled cell phones using the WML (Wireless Markup Language) format.

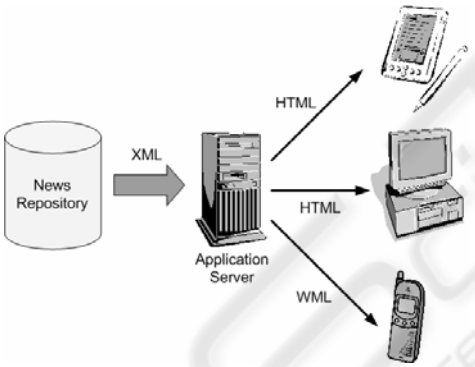


Figure 3: Data Provider

# 3 CONCLUSIONS

The rapid growth of the amount of information on Internet makes locating a specific subset a tedious and time-consuming task. Due to this, in the next years, automatic extraction and integration of web-based information will become more and more important. This paper proposed News Miner, an automatic tool for news extraction and integration based on XML language. It extracts relevant data from HTML pages using XSLT language. A weakness of this approach is that the extraction mechanism depends on a human input describing the HTML structure. This becomes an issue when the structure of web sources changes rapidly requiring frequent updates to the style sheets. In the future, we intend to develop a GUI for helping users to generate and maintain correct XSLT style sheets. In particular, it will allow the user simply to highlight the information that is to be extracted directly on the screen without having to write a single line of XSLT code. Extracting news from web services, based on protocols, such as SOAP (Simple Object Access Protocol) or XML-RPC (XML Remote Procedure Call), will provide relevant enhancements to the architecture.

# REFERENCES

Ashish N., Knoblock C., 1997. Wrapper Generation for Semi-structured Internet Sources. *Workshop on Management of Semistructured Data, Ventana Canyon Resort, Tucson, Arizona*.

Cooley R., Mobasher B., Srivastava J., 1997. Web Mining: Information and Pattern Discovery on the World Wide Web, *In ICTAI '97, 9th International Conference on Tools with Artificial Intelligence*.

DOM (Document Object Model) specifications – web site: http://www.w3.org/DOM

Etzioni O., 1996. The World Wide Web: quagmire or gold mine?, *In Communications of the ACM 39(11)*.

Kleinberg J. M., 1998. Authorative Sources in a Hyperlinked Enviroment, *In Proc. of the ACM-SIAM Symposium on Discrete Algorithms*.

Mobasher B., Jain N., Han E.-H., Srivastava J., 1997. Web Mining: Patterns from WWW Transactions. *Dept. Comput. Sci., Univ. Minnesota, Tech. Rep. TR96-050*.

Sillitti A., Vernazza T., Succi G., 2002. Service Oriented Programming: A New Paradigm of Software Reuse. *In Seventh International Conference on Software Reuse ICSR-7*.

Srivastava J., Cooley R., Deshpande M., Tan P., 2000. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, *In SIGKDD Explorations, Vol. 1, Issue 2, 2000.*

WML (Wireless Markup Language) specifications – web site: http://www.wapforum.org/what

XHTML (eXtensible HyperText Markup Language) specifications – web site: http://www.w3.org/MarkUp/

XPath (XML Path Language) specifications – web site: http://www.w3.org/TR/xpath

XSLT (eXtensible Stylesheet Language Transformation) specifications – web site: http://www.w3.org/Style/XSL/