The State of the Art in Web Mining

D

Tad Gonsalves
Department of Information and Communication Sciences, Sophia University, Japan

INTRODUCTION

Data Mining deals with extracting valuable and useful information and knowledge from large datasets (Hand et al., 2001). Three types of mining are well-known in the research mining community: data mining, web mining, and text mining. Data mining mainly deals with structured data organized in databases; text mining mainly deals with mining text. Web mining lies in between and copes with semi-structured data and/or unstructured data. Web mining calls for creative use of data mining and/or text mining techniques. Mining the web data is one of the most challenging tasks for the data mining researchers because the web is a huge warehouse of heterogeneous and semi-structured data.

Web mining is categorized into: Web content mining, Web usage mining and Web structure mining. Web content mining deals with the knowledge discovery, in which the main objects are the collections of text documents and, more recently, also the collections of multimedia documents. Web usage mining deals with the discovery of interesting patterns of user's usage of data on the web. Web structure mining deals with the analysis of the connection structure of a web site. Each of these categories may be further divided into several sub-categories. In practice, the three web mining tasks above could be used in isolation or combined in an application, especially in web content and structure mining since the web document might also contain links.

Kosala and Blockeel (2000) present a survey of web mining research for each of the three web mining categories presented above, and distinguish web mining as different from information retrieval and information extraction. They hold that web mining techniques are not the only tools to solve information overload problems either directly or indirectly. They claim that other techniques and works from different research areas, such as database, information retrieval, natural language processing could also be used.

This article introduces some of the state of the art applications in Web mining developed by the academia and industry. It introduces some of the highly successful Web mining applications such as e-commerce (data mining application in online business, e-search (web search), e-education (distance learning) and e-auction (online auction).

Finally, three areas, namely, Semantic Web Mining, Privacy Policy and Web Application Security are suggested where the current Web Mining technology need to further develop. The current applications collect a lot of data about the individual users to design and present a personalized page to the user and thereby improve the enterprise business. However, there is a danger of violating the users' privacy. This is one of the pressing issues the Web mining community should address. Other areas for future development in Web Mining are applications security and Semantic Web.

BACKGROUND

In Informatics, *data*, *information* and *knowledge* form a pyramid with *data* at the base, *information* in the middle and *knowledge* on top. *Data* refers to the facts which give a description of the world, *information* is data captured, while *knowledge* is our mental map or model of the world helping us to make informed decisions. The three are related by the act of processing – data can be processed into information and information in turn can be processed into knowledge.

One of the major problems of our data-ridden age is succinctly described by John Naisbett (1988): "We are drowning in information, but starving for knowledge." We can further extend this statement to include the fact that "We are drowning in data, but starving for information." Data mining - the science of extracting useful information (knowledge) from large data sets attempts to bridge the gaps among *data*, *information* and *knowledge*.

DOI: 10.4018/978-1-4666-5888-2.ch187

Copyright © 2015, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Conventional data-mining has been dedicated to the task of finding some knowledge patterns in data bases. Later it was extended to text mining. And now the stage is set for carrying the fully developed and mature data mining technology for harvesting the data and information spread across the World Wide Web. The major challenge is that data-mining tools and methods are tailored to work within the confines of highly structured and rigid information. The web on the other hand is a collection of semi-structured documents. This article describes the state-of-the-art of web mining technology which many business organizations employ to gather information and/or knowledge in order to utilize it in their best interest.

WEB MINING

Beginning with a working definition of Web mining, this section presents the peculiar characteristics of Web mining as opposed to data and text mining and presents the details of the three categories of Web mining, namely, content mining, structure mining and usage mining.

Characteristics of Web Mining

The term "Web Mining" was invented by Etzioni, as the use of data mining techniques to automatically discover and extract information from web documents and services (Etzioni, 1996). It is "the nontrivial process of identifying valid, previously unknown, and potentially useful patterns" in the Web data (Fayyad et al., 1996). Other authors define it as, "the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web (Romero, 2007; Liu, 2007) or as "the intelligent analysis of Web data (Song & Shepperd, 2006) or as the extension of the traditional data mining methodologies to Web mining (Zaiane et al., 1998; Cooley et al., 1999; Nasraoui et al., 2000; Srivastava et al. 2000).

Web mining is a challenging task, because the data is not arranged in a standard format ready for use. It has the following daunting characteristics (Liu & Chang, 2004):

- The amount of data/information on the Web is colossal and is still growing rapidly.
- Data of all types exist on the Web, e.g., structured tables, texts, multimedia, etc. Besides, it is semi-structured due to the nested structure of the HTML code.
- Information on the Web is heterogeneous. Multiple Web pages may present the same or similar information using completely different formats or syntaxes, which makes integration of information a challenging task.
- Much of the Web information is redundant. The same piece of information or its variations may appear in many pages or sites.
- The Web is noisy. A Web page typically contains a mixture of many kinds of information, e.g., main content, advertisements, navigation panels, copyright notices, etc. For a particular application only part of the information is useful, and the rest are noises.
- The Web is dynamic. Information on the Web changes constantly. Keeping up with the changes and monitoring the changes are important issues for many applications.

In general, Web mining tasks can be classified into three categories: Web content mining, Web structure mining and Web usage mining (Figure 1) (Markov et al., 2007).

Web Content Mining

Web content mining deals with the knowledge discovery, in which the main objects are the collections of text documents and, more recently, also the collections of multimedia documents. Accordingly, Web content mining can be divided into text mining (including text file, HTML document, etc.) and multimedia mining (including images, videos, audios, etc., which are embedded in or linked to the Web pages).

Although multimedia mining is recently drawing more interests, text mining is the most fundamental and important task as text is the primary source of information. Further, the main sub-categories of Web text

9 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/the-state-of-the-art-in-web-mining/112599

Related Content

Improved Secure Data Transfer Using Video Steganographic Technique
V. Lokeswara Reddy (2017). *International Journal of Rough Sets and Data Analysis (pp. 55-70).*
www.irma-international.org/article/improved-secure-data-transfer-using-video-steganographic-technique/182291

Federal Government Application of the Cloud Computing Application Integration Model
John P. Sahlin (2015). *Encyclopedia of Information Science and Technology, Third Edition (pp. 2735-2744).*
www.irma-international.org/chapter/federal-government-application-of-the-cloud-computing-application-integration-model/112692

A Hierarchical Hadoop Framework to Handle Big Data in Geo-Distributed Computing Environments
Orazio Tomarchio, Giuseppe Di Modica, Marco Cavalloand Carmelo Polito (2018). *International Journal of Information Technologies and Systems Approach (pp. 16-47).*
www.irma-international.org/article/a-hierarchical-hadoop-framework-to-handle-big-data-in-geo-distributed-computing-environments/193591

Sense and Boundaries of Computer Simulations
Georgios O. Papadopoulosand Apostolos Syropoulos (2021). *Encyclopedia of Information Science and Technology, Fifth Edition (pp. 155-163).*
www.irma-international.org/chapter/sense-and-boundaries-of-computer-simulations/260183

Medical Social Networks, Epidemiology and Health Systems
Patrícia C. T. Gonçalves, Ana S. Moura, M. Natália D. S. Cordeiroand Pedro Campos (2021). *Encyclopedia of Information Science and Technology, Fifth Edition (pp. 1827-1838).*
www.irma-international.org/chapter/medical-social-networks-epidemiology-and-health-systems/260310