Chapter XVII

A Java Technology Based Distributed Software Architecture for Web Usage Mining

Juan M. Hernansáez
University of Murcia, Spain

Juan A. Botía
University of Murcia, Spain

Antonio F.G. Skarmeta
University of Murcia, Spain

ABSTRACT

In this chapter we focus on the three approaches that seem to be the most successful ones in the Web usage mining area: clustering, association rules and sequential patterns. We will discuss some techniques from each one of these approaches, and then we will show the benefits of using METALA (a META-Learning Architecture) as an integrating tool not only for the discussed Web usage mining techniques, but also for inductive learning algorithms. As we will show, this architecture can also be used to generate new theories and models that can be useful to provide new generic applications for several supervised and non-supervised learning paradigms. As a particular example of a Web usage mining application, we will report our work for a medium-sized commercial company, and we will discuss some interesting properties and conclusions that we have obtained from our reporting.
INTRODUCTION

When we face the challenge of data recovering from the Internet, or Web mining (WM), we can consider it from two main perspectives: Web resources mining and Web usage mining (WUM). Still, we can split the first one into Web content mining and Web structure mining. While the differences between these resource mining areas sometimes are not clear, WUM is more clearly defined, but it is not isolated either. It aims to describe the behaviour of the users who are surfing the Web. Many techniques and tools have been proposed, giving partial solutions to some of the WUM problems. All the proposed techniques and ideas are showing their importance in many areas, including Web content and Web structure mining.

One of the most accepted definitions for Web usage mining is that it is “the application of data mining techniques to large Web data repositories in order to extract usage patterns” (Cooley, Tan & Srivastava, 1999). As we know, Web servers around the world record data about user interaction with the Web pages hosted in the Web servers. If we analyze the Web access logs of different Websites we can know more about the user behaviour and the Web structure, making easier the improvement of the design of the sites among many other applications.

Analyzing data from the access logs can help organizations and companies holding the Web servers to determine the ideal life cycle of their products, the customer needs, effectiveness of new launched products, and more. In short, WUM can support marketing strategies across products over some specific groups of users, and improve the presence of the organization by redesigning the Internet Websites. Even if the organization is based on intranet technologies, WUM can warn of a more effective infrastructure for the organization, and can also warn of improvements and faults of the workgroup communication channels.

This chapter is organized as follows: in the first section, we introduce the three approaches intended for solving the WUM problems and challenges: clustering, association rules and sequential patterns. We will focus on the techniques that we have integrated in our own architecture, having one technique of each approach. Then, in the second section, we present our software architecture for automated data analysis processes, called METALA. We will show how this architecture can be used to support the WUM techniques and models, as well as other data analysis methods like machine learning, and how we integrated some algorithms of each one of these three approaches.

As an example of a WUM application provided by METALA, in the third section we summarize our work for a medium-sized industrial company, which wanted to know more about the usage of its Website, and about the strengths and flaws of the site and how to improve it. From this work we have proposed some ideas to face the problems we found when applying the WUM techniques. Finally, in the fourth section we give our conclusions and discuss our future work.

WEB USAGE MINING TECHNIQUES

As we have already mentioned, there are three main Web usage mining approaches. They are clustering, association rules and sequential patterns. Most of the WUM techniques can be included in one of these approaches, and some others may be hybrid.
Theoretical Foundations of a CSCL Script in Persistent Virtual Worlds
According to the Contemporary Learning Theories and Models
[www.igi-global.com/chapter/theoretical-foundations-of-a-cscl-script-in-persistent-virtual-worlds-according-to-the-contemporary-learning-theories-and-models/137413?camid=4v1a](www.igi-global.com/chapter/theoretical-foundations-of-a-cscl-script-in-persistent-virtual-worlds-according-to-the-contemporary-learning-theories-and-models/137413?camid=4v1a)

Development and Testing of New E-Learning Hypermedia System
[www.igi-global.com/article/development-and-testing-of-new-e-learning-hypermedia-system/138292?camid=4v1a](www.igi-global.com/article/development-and-testing-of-new-e-learning-hypermedia-system/138292?camid=4v1a)