A Constraint Programming Approach for Web Log Mining

Amina Kemmar, LITIO, University of Oran 1, Ahmed Ben Bella, Oran, Algeria
Yahia Lebbah, LITIO, University of Oran 1, Ahmed Ben Bella, Oran, Algeria
Samir Loudni, GREYC (CNRS UMR 6072), University of Caen, Caen, France

ABSTRACT

Mining web access patterns consists in extracting knowledge from server log files. This problem is represented as a sequential pattern mining problem (SPM) which allows to extract patterns which are sequences of accesses that occur frequently in the web log file. There are in the literature many efficient algorithms to solve SPM (e.g., GSP, SPADE, PrefixSpan, WAP-tree, LAPIN, PLWAP). Despite the effectiveness of these methods, they do not allow to express and to handle new constraints defined on patterns, new implementations are required. Recently, many approaches based on constraint programming (CP) was proposed to solve SPM in a declarative and generic way. Since no CP-based approach was applied for mining web access patterns, the authors introduce in this paper an efficient CP-based approach for solving the web log mining problem. They bring back the problem of web log mining to SPM within a CP environment which enables to handle various constraints. Experimental results on non-trivial web log mining problems show the effectiveness of the authors’ CP-based mining approach.

KEYWORDS:
Constraint Programming, Constraint-Based Mining, Sequential Patterns, Web Logs, Web Mining

1. INTRODUCTION

Sequential pattern mining (SPM) is a well-known data mining technique introduced in (Srikant & Agrawal, 1995) to find regularities in a database of sequences. With the increasing amount of data available on the World Wide Web (WWW), discovering and analyzing useful information from the WWW becomes an essential task (Blockeel & Kosala, 2000). The task of Web Mining consists to extract an interesting and potentially meaningful patterns based on activities related to the WWW (Kohavi, 2002). There are roughly three knowledge discovery domains that pertain to web mining (Cooley, 1997): web content mining, web structure mining and web usage mining. In Web Usage Mining (WUM), also known as web access, the mining process aims at extracting knowledge from server log files collected when users access Web servers over a period of time (Kohavi, 2002).

Mining web access patterns from Web logs can provide useful information for server performance enhancements, restructuring a web site, and direct marketing in e-commerce (Etzioni & Oren, 1996). For an e-commerce company, this may help detecting future customers likely to make a large number of purchases, or predicting which online visitors will click on what commercials or banners, based on observation of prior visitors, who have behaved both positively or negatively to the advertisement banner (Ezeife & Yi, 2005).

There are already in the literature many efficient algorithms to extract sequential patterns (e.g. GSP (Agrawal & Srikant, 1996), SPADE (Zaki, 2001), PrefixSpan (Pei, Han, Mortazavi-Asl, Pinto, Dayal, & Hsu, 2001), web sequential patterns on web log data (e.g. WAP-tree (Jian Pei J. H.-A., 2000)),

DOI: 10.4018/IJITWE.2016100102

Copyright © 2016, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
PLWAP (Ezeife & Yi, 2005)), closed sequential patterns (e.g. CloSpan (Yan, Han, & Afshar, 2003), BIDE (Han & Wang, 2004)) or sequential patterns satisfying regular expressions (e.g. SPIRIT (Minos, Garofalakis, & Rastogi, 2002)). All the above methods, though efficient, they suffer from two major problems. First, they tackle particular classes of constraints (i.e. monotonic and anti-monotonic ones (see Section 5.3)) by using devoted techniques. However, several practical constraints required in data mining tasks, such as regular expression, gap, aggregates, do not fit into these classes. Second, they suffer from a lack of genericity to handle simultaneously different types of constraints in a simple way. Indeed, new constraints have to be hand-coded and their combinations often require the revision of the original algorithm to take into account the new constraints.

To address the above obstacles, many approaches based on constraint programming (CP) was proposed to solve the sequential pattern mining problem in a declarative and generic way (Coquery, Jabbour, Saïs, & Salhi, 2012; Métivier, Loudni, & Charnois, 2013; Kemmar, et al., 2014). The major benefit of these proposals is their flexibility to handle various constraints simultaneously. The user can simply model the problem and then express the different constraints to be satisfied by the extracted patterns without having to develop new algorithms from scratch. But unfortunately, these CP approaches are not enough efficient to handle large sequence datasets, which is explained by using a complex model based on reified constraints. More recently, the global constraint Prefix-Projection has been proposed for SPM in order to handle efficiently the frequency constraint and remedies to this drawback (Kemmar, Loudni, Lebbah, Boizumault, & Charnois, 2015). Prefix-Projection allows to encode efficiently the subsequence and frequency constraints and its filtering relies on the principle of projected databases (Pei, Han, Mortazavi-Asl, Pinto, Dayal, & Hsu, 2001).

Since no CP-based approach has been used for web log mining, we propose, in this paper, mining sequential patterns with Prefix-Projection global constraint in web log data. We show how its concise encoding allows for a straightforward implementation of the frequency constraint and the subsequence relation and other constraints on patterns such as size, item, regular expression and the simultaneous combination of them. Finally, experiments show that our approach competes well with specialized methods on large datasets and achieves scalability on web log files which is an important issue for CP approaches.

The paper is organized as follows. Section 2 first discusses some related works on Web logs mining and gives a critical review of specialized methods and CP approaches for SPM. Section 3 focuses on the preprocessing task. Section 4 states formally the problem of web log mining. Section 5 introduces our constraint programming model for web log mining and describes the global constraint Prefix-Projection. Some experimental results are discussed in Section 6. Finally, we conclude drawing some perspectives.

2. RELATED WORKS

First, we discuss the most important specialized methods for mining sequential patterns, including those which are devoted for mining web log data. Then, we give a critical review of CP approaches. The problem of mining sequential patterns was introduced in the context of basket data analysis. In this case, a transaction is represented by the set of items bought together in different interval times. In order to analyze the costumer behavior, we have to extract sequences which are the most frequent. GSP (Agrawal & Srikant, 1996) was the first algorithm which is based on a generate-and-test approach. This approach requires two steps: (1) generating the candidate and then, (2) computing its support. The generate-and-test methods allow to generate a huge number of candidates because no pruning
Data Cleansing and Validation for Multiple Site Link Structure Analysis
www.igi-global.com/chapter/data-cleansing-validation-multiple-site/31140?camid=4v1a

A Case Study: Creating and Designing a Bilingual Resource Web Site for Somali Immigrants
www.igi-global.com/chapter/case-study-creating-designing-bilingual/4951?camid=4v1a