# Mining Unexpected Web Usage Behaviors

Dong (Haoyuan) Li[1], Anne Laurent[2], and Pascal Poncelet[1]

[1] LGI2P - École des Mines d'Alès, Parc Scientifique G. Besse, 30035 Nîmes, France
{Haoyuan.Li,Pascal.Poncelet}@ema.fr
[2] LIRMM - Université Montpellier II, 161 rue Ada, 34392 Montpellier, France
laurent@lirmm.fr

**Abstract.** Recently, the applications of Web usage mining are more and more concentrated on finding valuable user behaviors from Web navigation record data, where the sequential pattern model has been well adapted. However with the growth of the explored user behaviors, the decision makers will be more and more interested in unexpected behaviors, but not only in those already confirmed. In this paper, we present our approach USER, that finds unexpected sequences and implication rules from sequential data with user defined beliefs, for mining unexpected behaviors from Web access logs. Our experiments with the belief bases constructed from explored user behaviors show that our approach is useful to extract unexpected behaviors for improving the Web site structures and user experiences.

## 1 Introduction

Recently, the applications of Web usage mining are more and more concentrated on finding valuable user behaviors from Web navigation record data (also known as Web access logs). A great deal of research work has been performed on porting data mining technologies to the Web usage analysis, in order to improve the personalization, the recommendation, and even the effectiveness of Web sites [1,2,3,4,5,6,7,8,9,10] by exploring the question: *what resources are frequently visited by whom during which periods?*

Among existing technologies, *sequential pattern* mining [11] has been well adapted to answer the above question [4,6,7,8,9]. All those sequential patterns extracted from Web access logs are typically the relationships like "on the Web site of customer support forum, 40% of users visited the `TopicList` page, then the `Search` page, then the `Login` page, and then the `PostTopic` page", or like "in the online store, 10% of customers visited the `notebook cases` page after having added a `notebook computer` to the `shopping cart`". This kind of relationships reflect the most general and reasonable user behaviors during Web navigations, however it become less important once we interpreted them as domain knowledge. When we regularly perform sequential pattern based Web usage mining on access logs, with the growth of the explored user behaviors, the decision makers will be more and more interested in exploring unexpected user behaviors that contradict existing knowledge, but not only in those the already confirmed.

P. Perner (Ed.): ICDM 2008, LNAI 5077, pp. 283–297, 2008.

In this paper, we focus on finding *unexpected behaviors* (that contradict the explored user behaviors) from Web access logs within the context of domain knowledge (that corresponds to the explored user behaviors). To illustrate our goal, let us consider an online news Web site, where the latest news are listed on the static home page `index.html` by categories. The latest previous news can be visited from static category index pages like `cat1.html`, `cat2.html`, etc., and all news can be visited from server side script page `listnews.php` by specifying the category, like `listnews.php?cat=1&page=3`. The server side script page `readnews.php` provides the detail of a specified news identified by `news`, like `readnews.php?news=20080114-002`. Assume that (1) 60% of users visit `index.html`, then various `readnews.php`, then `cat1.html`, then various `readnews.php`, then `listnews.php`, then various `readnews.php`, and then other categories and various `readnews.php`, etc.; (2) 10% of users visit `index.html`, then `cat5.html`, then various `readnews.php`; (3) 8% of users visit `readnews.php` only once; (4) 0.005% of users visit a large number of `readnews.php` only. From traditional sequential pattern mining approaches, we may find the most general user behaviors described in (1) with a suitable minimum support threshold, but it is quite hard to find the behaviors described in (2), (3) and (4) because:

1. Most existing sequential pattern mining approaches do not consider the missing elements, neither the semantic contradictions between elements (e.g. between `cat1.html` and `cat5.html`) in a sequence. The constraint based approaches like SPIRIT [12] may find the sequences of (2) and (3), but the main drawback is that we cannot find all sequences like the one described in (2) by saying "categories contradicting `cat1.html`", but will have to indicate `cat2.html`, `cat3.html`, etc. exactly, since the constraint **not** `cat1.html` implies all pages different to `cat1.html`.
2. According to the model of sequential patterns, the sequences representing (2), (3) and (4) are *contained* in the sequences representing (1). Existing approaches that distinguish the support value of each frequent sequence (instead of maximal frequent sequence), like the *closed sequential pattern* [13], may find the existence of (2), (3) and (4) by computing and comparing the support values, but it is also difficult to indicate them.

The rest of this paper is organized as follows. Section 2 presents the application of our approach USER (Mining Unexpected SEquential Rules) for finding unexpected behaviors from Web access log files. In Sect. 3 we show our experimental results. We introduce the related work in Sect. 4. The conclusion is listed in Sect. 5.

## 2 Finding Unexpected Behaviors from Web Access Logs

In this section, we present the application of our approach USER for finding unexpected Web usage behaviors. We first propose a formal definition of the session sequence contained in Web access logs, then we detail our approach USER, that finds unexpected sequences and implication rules with user defined