Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests

Haibin Liu, Vlado Kešelj *

Faculty of Computer Science, Dalhousie University, 6050 University Ave, Halifax, NS, Canada B3H 1W5

Abstract

We present a study of the automatic classification of web user navigation patterns and propose a novel approach to classifying user navigation patterns and predicting users' future requests. The approach is based on the combined mining of Web server logs and the contents of the retrieved web pages. The textual content of web pages is captured through extraction of character N-grams, which are combined with Web server log files to derive user navigation profiles. The approach is implemented as an experimental system, and its performance is evaluated based on two tasks: classification and prediction. The system achieves the classification accuracy of nearly 70% and the prediction accuracy of about 65%, which is about 20% higher than the classification accuracy by mining Web server logs alone. This approach may be used to facilitate better web personalization and website organization.
© 2006 Elsevier B.V. All rights reserved.

Keywords: Web usage mining; Web content mining; User navigation profiles; Classification; Prediction

1. Introduction

With the explosive growth of knowledge available on the World Wide Web, which lacks an integrated structure or schema, it becomes much more difficult for users to access relevant information efficiently. Meanwhile, the substantial increase in the number of websites presents a challenging task for webmasters to organize the contents of the websites to cater to the needs of users. Automatic classification of user navigation patterns provides a useful tool to solve these problems. The basis of our approach is extraction of navigation profiles that capture similar behaviour of site users. On one hand, the profiles can be used for predicting the navigation behaviour of current users, thus aiding in web personalization. On the other hand, webmasters can improve the design and organization of websites based on the acquired profiles.

---

* Corresponding author. Tel.: +1 902 494 2893; fax: +1 902 492 1517.
 E-mail address: vlado@cs.dal.ca (V. Kešelj).

From the user perspective, the classification of navigation patterns can enhance the quality of personalized web recommendations that aim to predict which web pages are more likely to be accessed next by current users. We estimate the best profile describing navigation behaviour of the current user, and find related unrequested web pages of great potential to be the next pages that the user wants to see. As the recommendations, the links of these pages will then be inserted into the currently requested page dynamically for display. This will help users access their favourite information efficiently. From the perspective of websites, the classification of navigation patterns can guide webmasters to organize the contents of sites. Instead of being arranged purely according to one view of the web site content, a site will be adjusted in terms of the desires of users. For instance, necessary links will be added between the web pages, which seemingly do not share the same topic, but were visited one after another by plenty of users. Also, pages which drew lots of clicks will be highlighted from their categories of topics, while pages which were not visited for a period of time will be moved or discarded. In fact, organizing websites by topics is both static and reactive. Since users' navigation patterns will be learned periodically, the change of their navigation interest can be captured regularly and then the site organization can be adjusted accordingly. This is a dynamic and proactive way of managing websites. As a result, the passing visitors will be enticed to become consumers or users of the site while current users are willing to remain loyal to the site.

Recently, web usage mining techniques have been widely applied for discovering interesting and frequent user navigation patterns from Web server logs. Sequential pattern mining [1], association rule mining [2,3] and clustering [4,5] discover different access patterns from web logs that can be modeled and used to offer a personalized and proactive view of the web services to users. At the same time, web content mining approaches have also been investigated and implemented for extracting knowledge from the contents of websites. For example, the classification of web pages is a typical application of content mining techniques [6].

While many results were reported in the web usage and content mining separately, few efforts were made to integrate these two aspects for a more effective classification of user navigation patterns. A project aiming at extracting navigation behaviour models of a site's visitors was introduced in [7]. In the project, two classification-type experiments were implemented, but the results were not very good with lower classification accuracy. One reason for such results discussed by the authors was they did not exploit content mining techniques. However, the contents of accessed pages may reveal topics related to visitors' profiles, which can improve the classification accuracy. In fact, it is often observable that web pages a visitor requested during his navigation are correlated and can be bound to a specific content. Thus, it is more likely that this type of content could be accessed in the further extension of this navigation. We therefore hypothesize that looking into web page contents will better capture site users' interests and increase the classification accuracy of user navigation patterns.

Inspired by the work of Baglioni et al. [7], we propose an experimental system to investigate whether associating a content mining approach with regular web usage mining could result in a more accurate classification of user navigation patterns, and consequently lead to a more accurate prediction of users' future requests. In this paper, we used character *N*-grams to represent the contents of web pages, and combined them with user navigation patterns by building user navigation profiles composed of a collection of *N*-grams. This character *N*-gram-based approach is a new way to integrate web usage mining and web content mining. Meanwhile, we tried different ways when building user profiles and attempted different parameters when experimenting on classification and prediction. Furthermore, we found that some parameters could influence the classification and prediction results. The existence of the optimal parameters reveals a clue on how to build desired user profiles and also becomes a guide for further experiments. We evaluated the performance of the experimental system using two defined measures: *classification accuracy* and *prediction accuracy*. We concluded that our system can achieve good experimental results although it is only a proof-of-concept prototype of the idea of combining both web usage and content mining.

The rest of this paper is organized as follows: In Section 2, we review recent research advances in both web usage and content mining. Section 3 describes the architecture of the experimental system proposed for classifying user navigation patterns and predicting users' future requests. The approaches and algorithms applied in the system are also explicated. The overall performance of the system is evaluated in Section 4. Finally, Section 5 summarizes the paper and introduces future work.