A Hybrid Approach to Web Usage Mining

Søren E. Jespersen, Jesper Thorhauge, and Torben Bach Pedersen

Department of Computer Science, Aalborg University
{sej,jespert,tbp}@cs.auc.dk

Abstract. With the large number of companies using the Internet to distribute and collect information, knowledge discovery on the web has become an important research area. Web usage mining, which is the main topic of this paper, focuses on knowledge discovery from the clicks in the web log for a given site (the so-called click-stream), especially on analysis of *sequences* of clicks. Existing techniques for analyzing click sequences have different drawbacks, i.e., either huge storage requirements, excessive I/O cost, or scalability problems when additional information is introduced into the analysis.
In this paper we present a new *hybrid* approach for analyzing click sequences that aims to overcome these drawbacks. The approach is based on a novel combination of existing approaches, more specifically the Hypertext Probabilistic Grammar (HPG) and Click Fact Table approaches. The approach allows for additional information, e.g., user demographics, to be included in the analysis without introducing performance problems. The development is driven by experiences gained from industry collaboration. A prototype has been implemented and experiments are presented that show that the hybrid approach performs well compared to the existing approaches. This is especially true when mining sessions containing clicks with certain characteristics, i.e., when constraints are introduced. The approach is not limited to web log analysis, but can also be used for general sequence mining tasks.

1 Introduction

With the large number of companies using the Internet to distribute and collect information, knowledge discovery on the web, or *web mining* has become an important research area. Web mining can be divided into three areas, namely *web content mining*, *web structure mining* and *web usage mining* (also called web log mining) [8]. Web content mining focuses on discovery of information stored on the Internet, i.e., the various search engines. Web structure mining can be used when improving the structural design of a website. Web usage mining, the main topic of this paper, focuses on knowledge discovery from the usage of individual web sites. Web usage mining is mainly based on the activities recorded in the *web log*, the log file written by the web server recording individual requests made to the server. An important notion in a web log is the existence of *user sessions*. A user session is a sequence of requests from a single user within a certain time window. Of particular interest is the discovery of frequently performed *sequences* of actions by web user, i.e., frequent sequences of visited web pages.

The work presented in this paper has been motivated by collaboration with the Zenaria company [21]. For more on this collaboration, please consult the full paper[10].

Y. Kambayashi, W. Winiwarter, M. Arikawa (Eds.): DaWaK 2002, LNCS 2454, pp. 73–82, 2002.

Much work has been performed on extracting various pattern information from web logs and the application of the discovered knowledge range from improving the design and structure of a web site to enabling companies to provide more targeted marketing. One line of work features techniques for working directly on the log file [8,9]. Another line of work concentrates on creating aggregated structures of the information in the web log [13,16]. The Hypertext Probabilistic Grammar (HPG) model [3,4], utilizing the theory of grammars, is such an aggregated structure. Yet another line of work focuses on using database technology in the clickstream analysis [1,7], building so-called "data webhouses" [12]. Several database schemas have been suggested, e.g. the click fact star schema where the individual click is the primary fact[12]. Several commercial tools for analyzing web logs exist [15,19,20], but their focus is mostly on statistical measures, e.g., most frequently visited pages and they provide only limited facilities for clickstream analysis. Finally, a prominent line of work focuses on mining *sequential patterns* in general sequence databases [2,13,14,17]. However, all the mentioned approaches have inherent weaknesses in that they either have huge storage requirements, slow performance due to many scans over the data, or problems when additional information, e.g., user demographics, are introduced into the analysis.

In this paper we present a new *hybrid* approach for analyzing click sequences that aims to overcome these drawbacks. The approach is based on a novel combination of existing approaches, more specifically the HPG and Click Fact Table [12] approaches. The new approach attempts to utilize the quick access and the flexibility with respect to additional information of the click fact table, and the capability of the HPG model to quickly mine rules from large amounts of data. Specialized information is extracted from the click fact schema and presented using the HPG model. The approach allows for additional information, e.g., user demographics, to be included in the analysis without introducing performance problems. A prototype has been implemented and experiments are presented that show that the hybrid approach performs very well compared to the existing approaches. This is especially true when mining sessions containing clicks with certain characteristics, i.e., when constraints are introduced. The approach is not limited to web log analysis, but can also be used for general sequence mining tasks.

We believe this paper to be the first to present an approach for mining frequent sequences in web logs that at the same time provides small storage requirements, very fast rule mining performance, and the ability to introduce additional information into the analysis with only a small performance penalty. This is done by exploiting existing data warehouse technology.

The paper is organized as follows. Section 2 describes the techniques on which we base our new hybrid approach. Section 3 describes the hybrid approach in detail. Section 4 describes the prototype implementation. Section 5 examines the performance of the hybrid approach. Section 6 concludes and points to future work.

2 Background

This section briefly describes the approaches underlying our hybrid approach. For a more in-depth discussion of the various technologies, please consult the full paper[10].