Chapter 4

Mining Frequent Generalized Patterns for Web Personalization in the Presence of Taxonomies

Panagiotis Giannikopoulos
University of Peloponnese, Greece

Iraklis Varlamis
Harokopio University of Athens, Greece

Magdalini Eirinaki
San Jose State University, USA

ABSTRACT

The Web is a continuously evolving environment, since its content is updated on a regular basis. As a result, the traditional usage-based approach to generate recommendations that takes as input the navigation paths recorded on the Web page level, is not as effective. Moreover, most of the content available online is either explicitly or implicitly characterized by a set of categories organized in a taxonomy, allowing the page-level navigation patterns to be generalized to a higher, aggregate level.

In this direction, the authors present the Frequent Generalized Pattern (FGP) algorithm. FGP takes as input the transaction data and a hierarchy of categories and produces generalized association rules that contain transaction items and/or item categories. The results can be used to generate association rules and subsequently recommendations for the users. The algorithm can be applied to the log files of a typical Web site; however, it can be more helpful in a Web 2.0 application, such as a feed aggregator or a digital library mediator, where content is semantically annotated and the taxonomic nature is more complex, requiring us to extend FGP in a version called FGP+. The authors experimentally evaluate both algorithms using Web log data collected from a newspaper Web site.

DOI: 10.4018/978-1-61350-474-1.ch004
INTRODUCTION

The role of recommendations is very important in everyday transactions. When buying a product, or reading a newspaper article, one would like to have recommendations on related items. To achieve this, recommendation engines first build a predictive model, by discovering itemsets or item sequences with high support among users. Recommendations are subsequently generated by matching new transaction patterns to the predictive model. Most current approaches in Web personalization consider that a Web site consists of a finite number of Web pages and build their predictive models based on this assumption (Mobasher, 2007). The Web, however, is a continuously evolving environment and this assumption does no longer hold. News portals are typical examples of this situation since they update their content on a regular basis. As a result, the traditional usage-based approach that takes as input the navigation paths recorded on the Web page level is not as effective. Since most predictive models are based on frequent itemsets, the more recent a page is, the more difficult it is to become part of the recommendation set; at the same time, such pages are more likely to be of interest for the average user. This problem can be addressed by generalizing the page-level navigation patterns to a higher, aggregate level (Eirinaki et. al. 2003; Mobasher, 2007).

In this work, we present the FGP algorithm, to address the aforementioned problem. The FGP algorithm is in essence the result of the modification and combination of two algorithms that have been proposed in different contexts. The first one, FP-Growth (Han et. al. 2004), is given a database of user transactions that comprise one or more unordered items (itemsets) and a minimum support threshold. The algorithm processes the transaction database and mines the complete set of frequent itemsets (whose frequency surpasses the threshold). FP-Growth considers the support of each item in the set to be equal to one. In this work, we extend the algorithm so that it assigns different weights to every item in the set depending on its importance in the transaction. We should note that the FP-Growth algorithm does not consider any relation between items in the database. This, however, is not the case in the Web, where items in a Web site are (conceptually) hierarchically organized. This intrinsic characteristic of the Web can be tackled by the second algorithm, GP-Close (Jiang and Tan, 2006; Jiang et. al., 2007). GP-Close considers a hierarchical organization of all items in the transaction database and uses this information to produce generalized patterns. The two algorithms are very efficient and solve many of the problems of pattern mining, such as the costly generation of candidate sets and the over-generalization of rules.

The FGP algorithm works efficiently in the case of Web sites that have a well-defined underlying hierarchy of topics, such as news portals. Many Web 2.0 sites, however, present a more complex underlying structure. For instance, feed aggregators summarize and present content that is collected from multiple sources. In such sites, the content is not necessarily classified into predefined categories (Inform 2007), being described by user-defined tags instead. This collaborative tagging process results into folksonomies (Voss 2007; P. Heymann and H. Garcia-Molina 2006) that differentiate from the traditional top-down taxonomies. The more complex structure of folksonomies, the use of plurals, the synonym polysemy and specificity of tagging raise new issues for the recommendation engines. In this context, we propose an extension of the FGP algorithm, named FGP+ that takes a more composite topic hierarchy as input, and supports multiple category assignments per topic.

In brief, the contributions of our work are outlined in what follows:

- We modify and combine the forces of FP-Growth and GP-Close in one efficient generalized pattern mining algorithm, named FGP which: