Expert Systems with Applications 37 (2010) 2911–2922

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Combination of Web page recommender systems

Murat Göksedef [1], Şule Gündüz-Öğüdücü *,[1]

Department of Computer Engineering, Istanbul Technical University, 34469 Istanbul, Turkey

ARTICLE INFO

Keywords:
Web usage mining
Web page recommendation
Hybrid recommender systems

ABSTRACT

With the rapid growth of the World Wide Web (www), finding useful information from the Internet has become a critical issue. Web recommender systems help users make decisions in this complex information space where the volume of information available to them is huge. Recently, a number of Web page recommender systems have been developed to extract the user behavior from the user's navigational path and predict the next request as s/he visits Web pages. However, each of these systems has its own merits and limitations. In this paper, we investigate a hybrid recommender system, which combines the results of several recommender techniques based on Web usage mining. We conduct a detailed comparative evaluation of how different combined methods and different recommendation techniques affect the prediction accuracy of the hybrid recommender. We then discuss the results in terms of using a hybrid recommender system instead of a single recommender model. Our results suggest that the hybrid recommender system is better in predicting the next request of a Web user.

1. Introduction

Most Web users complain about finding useful information on Web sites. Web page recommender systems predict the information needs of users and provide them with recommendations to facilitate their navigation. Given a user's current actions, the goal is to determine which Web pages will be accessed next. Many Web sites on Internet use Web page recommender systems to increase their usability and user satisfaction. Traditional methods for recommendation are based on Web usage and Web content mining techniques (Agrawal & Srikant, 1995; Cadez, Heckerman, Meek, Smyth, & White, 2003; Deshpande & Karypis, 2004; Nanopoulos, Katsaros, & Manolopoulos, 2001; Sarukkai, 2000).

In recent years, there has been an increasing interest in applying Web content mining techniques to build Web recommender systems. However, the Web content mining techniques are unable to handle constantly changing Web sites, such as news sites, and dynamically created Web pages. Thus, using Web content mining techniques in a recommender model leads to update the model frequently. For this reason, in this work our aim is to derive a predictive model for Web pages which is based on Web usage mining techniques.

The performance of a recommender model depends on the structure of the Web site besides the specific technique that it uses. Furthermore, different users may have different navigation strategies. Thus, it could be difficult to estimate a single best model for recommendation. Each of the single methods has its advantages, but also limitations and disadvantages. Therefore, combining different methods to overcome disadvantages and limitations of a single method may improve the performance of recommenders. Hybrid recommender systems combine two or more techniques to improve recommender performance. Burke (2002) proposed hybrid recommenders for collaborative filtering that combine information across different sources. To date, most of the research on hybrid recommenders is on collaborative filtering approaches such as combining these approaches with content based approaches rather than combining multiple recommender techniques. For this reason, we especially focused in this paper on how the recommendation accuracy may be improved by combining different recommendation models based on the Web usage mining techniques.

Web page recommendation systems have been extensively explored in Web usage mining. However, the quality of recommendations and the user satisfaction with such systems are still not optimal. Recommender systems based on Web usage mining techniques have also strengths and weaknesses (Kazienko & Kolodziejski, 2006). Since different methods focus on different characteristics of Web users, they produce different prediction accuracies on the same data set. Combination of different methods may result in better accuracy (Burke, 2002). For example, consider two recommender models, model *A* and model *B*, applied to a Web site to predict three pages one of which could be the next request of a Web user surfing on this site. Both models use different methods to predict the next request of a user. Let two users, user 1 and user 2, are requesting p_1 and p_4, respectively, where p_i is a Web page on the Web site. Model *A* generates the same

* Corresponding author. Tel./fax: +90 212 285 3597.
E-mail addresses: goksedef@itu.edu.tr (M. Göksedef), sgunduz@itu.edu.tr (Ş. Gündüz-Öğüdücü).
[1] The authors are given in alphabetical order.

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.09.046

recommendation set, $RS_A = \{p_1, p_2, p_3\}$, for both users, where model B generates $RS_B = \{p_2, p_4, p_5\}$ for the same users. In this case, model A is successful for user 1 and model B is successful for user 2. However, if it is possible to combine these two models into one model to generate a recommendation set, for example $RS = \{p_1, p_4, p_2\}$, this combined model can generate correct recommendations for both users.

In this study, we analyze the performance of hybrid recommender models by combining the results of different recommender techniques using four different hybridization (combination) methods. We especially focus on answering the questions of whether using hybrid approaches increases the recommendation accuracy and what types of combination methods are likely to be successful in predicting the next request of users. We use the structure information of the Web site for ranking the pages in the recommendation set. For this purpose, we implement and compare different hybrid recommenders including some novel combinations. The different recommender techniques implemented in this work are called the modules of the hybrid recommender model. We conduct a detailed comparative evaluation of different recommender techniques which can be used as the modules of the hybrid recommender system. For this purpose, four recommender techniques which use different data mining approaches based on different characteristics of user sessions are implemented (Cadez et al., 2003; Gündüz & Özsu, 2003; Mobasher, Dai, Luo, & Nakagawa, 2001, 2002a). We combine the modules by using four different combination methods. We propose a new combination method using Google's PageRank (Brin & Page, 1998) and HITS algorithms (Kleinberg, 1999) that make use of the topological structure of the Web sites. We also examine whether it is worthwhile to use a hybrid recommender instead of a single recommender model.

The hybrid recommender is implemented as an experimental system and its performance is evaluated based on the correct prediction of the next request of a user, namely Hit-Ratio. Our detailed experimental results show that when choosing appropriate combination methods and modules, hybrid approaches achieve a better prediction accuracy.

The rest of the paper is organized as follows: In Section 2, we introduce the related work. Section 3 describes the architecture of the experimental system designed for predicting the next request of users by a hybrid recommender. In Section 4, we introduce the distinct techniques which are integrated to build a hybrid recommendation model. The methods for combining different recommender techniques are explicated in Section 5. Section 6 provides detailed experimental results. Finally, in Section 7 we conclude and discuss future work.

2. Related work

Various Web usage mining techniques have been used to develop efficient and effective recommendation systems. One of the most successful and widely used technologies for building recommendation systems is Collaborative Filtering (CF) (Resnick & Varian, 1997). CF techniques predict the utility of items of an active user by matching, in real-time, the active user's preferences against similar records (nearest neighbors) obtained by the system over time from other users (Breese, Heckerman, & Kadie, 1998). A shortcoming of these approaches is that it becomes hard to maintain the prediction accuracy in a reasonable range while handling the large number of items in order to decrease the on-line prediction cost. Some hybrid approaches are proposed to handle these problems, which combine aspects of both pattern discovery methods and CF. O'Conner and Herlocker (1999) use existing data partitioning and clustering algorithms to partition the set of items based on user rating data. Predictions of items are then computed independently within each partition.

Another example to hybrid approaches is content-based systems, which work by comparing text descriptions or other representations associated with an item. Balabonović and Shoham (1995) describe a system that helps users to discover new and interesting sites that are of interest to them. As an alternative to the methods discussed above, the system proposed by Pazzani (1999) is designed to create a framework. In this work three major recommendation approaches including content based (using the product attributes), demographic filtering (using the customer attributes) and CF techniques are combined. Furthermore, hybrid approaches, that utilize content information, are proposed by Burke (2002) to overcome some of the shortcomings of CF such as the cold start problems for new users and new products. In this work six different hybridization methods are surveyed and implemented.

There have been attempts to combine content, structure and usage information to generate hybrid systems. One recent work in this area is Li and Zaïane's work (Li & Zaïane, 2004). They propose an algorithm that combines and utilizes of usage, content and structure data. In Nakagawa and Mobasher (2003), a recommender system is proposed that adopts a clustering technique to obtain both the site usage and site content profiles. In this work, the authors use association rule mining and sequential pattern mining to generate navigational patterns of Web users. A switching (Burke, 2002) hybridization method is used to integrate the navigational patterns of Web users in order to generate a recommendation set. In Göksedef and Gündüz-Öğüdücü (2007) a recommendation model, which is called consensus recommender, is developed. In the study, several recommendation models based on Web usage mining techniques are integrated. Results of the study show that consensus model achieves a better prediction accuracy compared to it's individual components.

3. System design

In this work, we have designed an experimental system to assist in our examination of whether combining multiple recommender techniques increases the prediction accuracy of a recommendation system. Fig. 1 depicts the overall process that we consider in our system. As with most recommender systems, ours is composed of two parts: an off-line part and an on-line part. The off-line part has two components, namely data preparation and pattern extraction. The on-line part also consists of two components: recommendation engine and hybridization.

3.1. Off-line part

3.1.1. Data preparation

The task of this component is to prepare the raw Web access log data for mining the usage patterns. Generally, several preparation steps need to be performed. For our work, these include data cleaning, user identification and session identification. These preprocessing steps are the same for any Web usage mining problem and fundamental methods of them have been well studied in Cooley, Mobasher, and Srivastava (1999) Srivastava, Cooley, Deshpande, and Tan (2000). In the data cleaning step, first the irrelevant log entries with filename suffixes such as, gif, jpeg, GIF, JPEG, jpg, JPG are eliminated. Next, the URLs in the log file are normalized in order to determine same Web pages which are represented by syntactically different URLs. A common form for each page is chosen using a Web crawler (Mohr, Kimpton, Stack, & Ranitovic, 2004). Only links that point to the Web pages within the site are added to the list of pages to explore. Comparing the content of pages pro-