Expert Systems with Applications 39 (2012) 1381–1387

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Hybrid personalized recommender system using centering-bunching based clustering algorithm

Subhash K. Shinde *, Uday Kulkarni

Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, Navi Mumbai 400 614, India
Department of Computer Science and Engineering, SGGS Institute of Engineering and Technology, Nanded, India

ARTICLE INFO

Keywords:
Collaborative filtering
Centering-bunching based clustering
Web personalized recommender system

ABSTRACT

In recent years, there is overload of products information on world wide web. A personalized recommendation is an enabling mechanism to overcome information overload occurred when shopping in an Internet marketplace. This paper proposes a novel centering-bunching based clustering (CBBC) algorithm which is used for hybrid personalized recommender system (CBBCHPRS). The proposed system works in two phases. In the first phase, opinions from the users are collected in the form of user-item rating matrix. They are clustered offline using CBBC into predetermined number clusters and stored in a database for future recommendation. In the second phase, the recommendations are generated online for active user using similarity measures by choosing the clusters with good quality rating. This helps to get further effectiveness and quality of recommendations for the active users. The experimental results using Iris dataset show that the proposed CBBC performs better than *K*-means and new *K*-medodis algorithms. The performance of CBBCHPRS is evaluated using Jester database available on website of California University, Berkeley and compared with ants recommender system (ARS). The results obtained empirically demonstrate that the proposed CBBCHPRS performs superiorly and alleviates problems such as cold-start, first-rater and sparsity.

1. Introduction

Many e-commerce Web sites offer numerous services, so a product search could return an overwhelming set of options. Without system support, filtering irrelevant products, comparing alternatives, and selecting the best option can be difficult or impossible. The recommender systems are personalized applications that can address this problem and suggest products (movies, music, jokes, books, news, web pages) that suit the user's needs (Adomavicius & Tuzhilin, 2005). These systems add related information of items to the information flowing towards the user, as opposed to removing information items. Typically, a recommender system compares the user's profile to some reference characteristics, and seeks to predict the rating that a user would give to an item they had not yet considered. Recommender systems use collaborative filtering approaches or a combination of the collaborative and content based filtering approaches, although content-based recommender systems do exist (Koren, Bell, & Chris, 2009). The web personalized recommendation systems (WPRS) are recently applied to provide different type of customized information for their users. The WPRS are applied in many areas such as: web-browsing, information filtering, net-news or movie recommender and e-commerce. The central element of all recommender systems is the user model that contains knowledge about the individual preferences which determine his or her behavior in a complex environment of web-based system. The WPRS are characterized by cross-fertilization of various research fields such as: information retrieval, artificial intelligence, knowledge representation, discovery and data/web mining, computational learning and intelligent and adaptive agents. The alternating information environment that is combined of various users, their needs and contexts of use as well as different system platforms necessitates application of recommender systems. The ever increasing importance of the e-commerce in the global economy also increases the importance of the WPRS. They are developed by different domains such as personal agents and adaptive hypermedia. The personalized hypermedia application is defined as a hypermedia system that adapts: the content, structure, and/or presentation of the web objects to each individual user's model.

The remainder of this paper is organized as follows. The section 2 describes various types of input data that are used for the recommendation systems. The section 3 summarizes the different techniques of recommendation systems and their drawbacks. The proposed clustering based hybrid personalized recommender system is described in the section 4. The section 5 illustrates experimental setup of the proposed recommendation system. This

* Corresponding author. Tel.: +91 09221788066; fax: +91 022 27473196.
E-mail addresses: skshinde@rediffmail.com (S.K. Shinde), kulkurniuv@yahoo.com (U. Kulkarni).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.08.020

section also gives performance evaluation with the existing algorithms. Finally, the section 6 concludes the paper.

2. Input to the recommendation systems

The recommendation systems are being widely used in many applications such as Amazon.com, Net-flix.com etc.; to suggest products, services, and information of items to potential consumers. At the heart of recommendation technologies are the algorithms for making recommendations based on various types of input data. In e-commerce, most recommendation algorithms take the following three types of data as an input: product attributes, consumer attributes, and previous interactions between consumers and products (e.g., buying, rating, and catalog browsing). The input data types are summarized in the Table 1.

3. Personalized recommendation techniques

In the recent years web personalization has undergone through tremendous changes. The content (Allen, 1990; Kalles, Papagelis, & Zaliagis, 2003), collaborative (Hofmann, 2003) and hybrid (Balabanovic & Sholam, 1997) based filtering are three basic approaches used to design recommendation systems.

The content based filtering (Chun Zeng et al., 2002) relies on the content of an item that user has experienced before. The content based information filtering has proven to be effective in locating text, items that are relevant to the topic using techniques such as Boolean queries, vector space queries etc. However, content based filtering has some limitations. It is difficult to provide appropriate recommendation because all the information is selected and recommended based on the content. Moreover, the content based filtering leads to overspecialization i.e. it recommends all the related items instead of the particular item liked by the user. The collaborative-filtering (Ulrike & Daniel, 2006) aims to identify users who have relevant interests and preferences by calculating similarities and dissimilarities between their profiles. The idea behind this method is that to one's search the information collected by consulting the behavior of other users who shares similar interests and whose opinions can be trusted may be beneficial. The different techniques have been proposed for collaborative recommendation; such as correlation based method, semantic indexing etc. The collaborative filtering overcomes some of the limitations of the content based filtering. The system can suggest items to the user, based on the rating of items, instead of the content of the items which can improve the quality of recommendations. However, collaborative filtering has some drawbacks. The first drawback is that the coverage of rating could be very sparse thereby resulting in poor quality recommendation. In the case of the addition of new items into database, the system would not be able to recommend until that item is served to a substantial number of users known as cold-start. Secondly, when new users are added, the system must learn the user preferences from the rating of users, in order to make accurate recommendations. Moreover, these recommendation algorithms seem to be very extensive and grow non-linearly when the number of users and items in a database increase. The hybrid recommendation systems (Adomavicius & Tuzhilin, 2005; Shinde & Kulkarni, 2008; Shinde & Kulkarni, xxxx; Yu, Schwaighofer, Tresp, Xu, & Kriegel, 2004) combine content and collaborative based filtering to overcome these limitations. As stated below, there are different ways of combining content and collaborative based filtering (Cheung & Tsui, 2004).

i. Implementing these approaches separately and combining them for prediction.
ii. Incorporating some content based characteristics into collaborative approach and vice versa.
iii. Constructing a general unified model that incorporates both content and collaborative based characteristics.

The hybrid approach proposed in this paper extracts user's current browsing patterns using web usage mining, and forms a cluster of items with similar psychology to obtain implicit users rating for the recommended item.

4. Proposed CBBCHPRS

We have developed and tested the CBBCHPRS for Jester dataset available on website of California University, Berkeley. The system architecture has been partitioned into two main phases; offline and online. The Fig. 1 depicts the architecture of CBBCHPRS with its essential components.

The phase I is offline. It does the preprocessing and clustering. In this phase background data in the form of user-item rating matrix is collected and clustered using the proposed approach which is described in section 4.1.2. Once the clusters are obtained the cluster data along with their centroids are stored for future recommendations.

The phase II is online in which the recommendation takes place for the active user. Here, similarity and density of clusters are calculated for choosing best clusters for making recommendations. The rating quality of each item unrated by active user is computed in the chosen clusters. To generate the recommendations, clusters are further selected based on rating quality of an item. The recommendations are then made by computing the weighted average of the rating of items in the selected clusters.

The working of CBBCHPRS is described below in detail with the Jester dataset.

4.1. Preprocessing phase

4.1.1. Normalization of data

User-item rating taken from Jester dataset rated in the scale of −10 to +10 is normalized in the scale of 0 to 1, where 0 indicates that item is not rated by corresponding user. To facilitate the discussion, running example shown in the Table 2 is used, where U_1-U_{10} are the users and J_1-J_{10} are the items (jokes) rated or unrated by users. The last row of Table 2 gives ratings of the active user.

4.1.2. Centering-bunching based clustering

In the *K*-means, and new *K*-medodis (Hae-Sang & Chi-Hyuck Jun, 2009) clustering algorithm centroids are initially selected by

Table 1
Taxonomy of input data.

1	Demographic data	name, age, gender, profession, birth date, telephone, address, hobbies, salary, education, experience and so on
2	Rating data	rating scores such as discrete multilevel and continuous ratings; and based on latest comments such as best, good, bad, worse and so on.
3	Behavior pattern data	duration of browsing, click times, the links of webs; save, print, scroll, delete, open, close, refresh of webs; selection, edition, search, copy, paste, and so on.
4	Transaction data	purchasing date, purchase quantity, price, discounting and so on
5	Production data	for movies, jokes or music, actor or singer, topic, release time, price, brand and so on.