Expert Systems With Applications 89 (2017) 404–412

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Content-based filtering for recommendation systems using multiattribute networks

Jieun Son, Seoung Bum Kim*

Department of Industrial Management Engineering, Korea University, 145 Anam-Ro, Seoungbuk-Gu, Anam-dong, Seoul 136-713, South Korea

ARTICLE INFO

Article history:
Received 21 April 2017
Revised 1 August 2017
Accepted 2 August 2017
Available online 3 August 2017

Keywords:
Content-based filtering
Recommender system
Movie recommendation
Network analysis

ABSTRACT

Content-based filtering (CBF), one of the most successful recommendation techniques, is based on correlations between contents. CBF uses item information, represented as attributes, to calculate the similarities between items. In this study, we propose a novel CBF method that uses a multiattribute network to effectively reflect several attributes when calculating correlations to recommend items to users. In the network analysis, we measure the similarities between directly and indirectly linked items. Moreover, our proposed method employs centrality and clustering techniques to consider the mutual relationships among items, as well as determine the structural patterns of these interactions. This mechanism ensures that a variety of items are recommended to the user, which improves the performance. We compared the proposed approach with existing approaches using MovieLens data, and found that our approach outperformed existing methods in terms of accuracy and robustness. Our proposed method can address the sparsity problem and over-specialization problem that frequently affect recommender systems. Furthermore, the proposed method depends only on ratings data obtained from a user's own past information, and so it is not affected by the cold start problem.

1. Introduction

The development of the Internet and web technologies has expanded the range of items available in various areas such as entertainment, e-commerce, and education. However, it is becoming difficult to find suitable items that match the preferences of users. Thus, recommender systems are expected to guide users to items that might be of interest to them. Various approaches have been introduced to improve the recommendation performance of recommender systems (Adomavicius & Tuzhilin, 2005; Park, Kim, Choi, & Kim, 2012), where bestseller recommendation is one of the simplest mechanisms, based on sales frequencies. In addition, ratings data, purchase history records, user demographic information, social information, and product descriptions are used to improve customer satisfaction in personalized recommender systems (Al-Shamri, 2016; Nunes, 2012). Recommender systems are typically categorized into collaborative filtering (CF) and content-based filtering (CBF) systems. CF algorithms attempt to predict the utility of an item for a particular user based on the similarity to the user's past ratings (Mahmoud & John, 2015). These methods are used widely, but the main limitation of CF systems is that they are vulnerable to fraud or profile injection attacks, which might have significant negative effects on the robustness of such systems (Lee et al., 2012). In particular, the performance of CF degrades as the numbers of customers and products increase, and thus CF algorithms must be updated continually over time. Therefore, new approaches are required that can rapidly produce high-quality recommendations, even for very large-scale problems (Cai et al., 2014).

CBF algorithms recommend suitable items to users based on the descriptions of items and user preferences. Furthermore, CBF algorithms employ profile information or ratings solely for the active user, and thus they can generate accurate recommendations even if the number of ratings from other users is not sufficiently large. Nevertheless, CBF has several drawbacks. CBF cannot generate suitable suggestions if the content analyzed for an item does not contain appropriate information for categorization. To address this limitation, feature weighting (FW) has been proposed for CBF, which assigns different levels of importance to different features (Debnath, Ganguly, & Mitra, 2008). For example, when choosing a cellular phone, the price may be more important than its color. However, this approach tends to provide biased results to users when items are recommended by computing the similarity repeatedly based on only a few attributes, which is usually called the over-specialization problem. To address these problems, some studies have attempted to include ontological information in semantic analyses. However, the scalability and the sparsity problem may occur when enormous amounts of data are calculated using matrix-based approaches (Di Noia, Mirizzi, Romito, Zanker (2012)).

* Corresponding author.
E-mail addresses: jieunson@korea.ac.kr (J. Son), sbkim1@korea.ac.kr (S.B. Kim).

http://dx.doi.org/10.1016/j.eswa.2017.08.008
0957-4174/

In the present study, we propose a type of CBF that uses a multiattribute network (MN), which comprises entire attribute information for different items. Many possible attributes of CBF can play significant roles in determining the quality of the recommendation results, because they may provide sufficient information for measuring sophisticated similarities. In the network analysis, we measure the similarities between directly and indirectly linked items. Moreover, centrality techniques in the network analysis can simultaneously consider the mutual relationships among items that are indirectly connected as well as determining the structural patterns of these interactions. This approach can address the sparsity problem and the over-specialization problem that frequently affect recommender systems. Furthermore, our proposed method depends only on ratings data obtained from a user's own past information, and so it is not affected by the cold start problem that is prevalent in CF (Lika, Kolomvatsos, & Hadjiefthymiades, 2014).

The remainder of this paper is organized as follows. In the related work section, we review CBF-based recommendation approaches. The proposed method section presents the proposed recommender system. The experimental evaluation section presents the experimental results and discussion. In the final section, we present our conclusions.

2. Related work

2.1. Multiattribute recommender systems

2.1.1. Pure CBF

CBF attempts to recommend items similar to those that a given user has liked in the past. The basic process is performed by matching user preferences with item attributes. Therefore, these systems require appropriate techniques for representing items and determining user preferences, as well as strategies for comparing user preferences with item representations (Choi, Kang, & Jeon, 2006; Lops, De Gemmis, & Semeraro, 2011).

Various machine learning techniques have been applied to CBF, including decision trees, K-means, neural networks, and naïve Bayes. For example, the basic concept employed by a naïve-Bayes classifier aims to determine whether an item is preferable by examining attribute information (Lew, Sebe, Djeraba, & Jain, 2006; Li, Lu, & Xuefeng, 2005). This method is used to estimate the probability that an item belongs to a class C_i The rating prediction is computed using the following probability function:

$$P(C_i|X) = \prod_{k=1}^{n} P(x_k|C_i), \tag{1}$$

where each item instance X is described by a conjunction of item attribute values $\langle x_1, x_2, \ldots x_k \rangle$. However, CBF cannot provide suitable recommendation results if the content does not contain sufficient information for classifying items. In some cases, domain knowledge or an ontology is required to determine the attributes that play important roles in recommendation.

2.1.2. Feature weighting

In CBF, feature weighting (FW) assigns different levels of importance to different features (Debnath et al., 2008), where the weight values obtained from a social network graph are used for predicting user preferences. The similarity S between objects O_i and O_j is calculated as

$$S(O_i, O_j) = \omega_1 f(A_{1i}, A_{1j}) + \omega_2 f(A_{2i}, A_{2j}) + \cdots + \omega_n f(A_{ni}, A_{nj}), \tag{2}$$

where ω_n is the weight for attribute A_n between object O_i and O_j, and f depends on the type of attribute. However, FW algorithms constitute hybrid recommender systems, because they require preference information about other users. Thus, they cannot be built on the basis of the ratings provided by "active users."

2.2. Semantic analysis

Some studies have aimed to improve the recommendation performance by including ontological information. Di Noia et al. (2012) demonstrated how the linked open data (LOD) cloud can be used as the main information source for a semantic CBF, where this method employs a three-dimensional vector space model comprising movies, movie properties, and the values of properties extracted from DBpedia and LinkedMDB. Semantic analysis using LOD is more suitable in comparison with existing methods. However, although the content analyzed for items contains sufficient information, the scalability problem still needs to be resolved. Furthermore, when a large number of attributes are included, sparse matrix issues can frequently occur.

2.3. Network analysis

Networks are patterns in the relationships that connect objects. A network analysis involves exploring, describing, and understanding numerous relational and structural aspects of networks. Important studies of network analyses have been conducted in an extremely diverse range of fields, including sociology, psychology, business, computer science, mathematics, and statistics (Anderson & Vongpanitlerd, 2006). The nodes in a network represent objects, and the links denote the relationships or flows between nodes. The distinguishing feature of a network analysis compared with more traditional connectivity measures is that it explicitly quantifies how the connectivity varies according to the characteristics of individuals. One of its advantages is that very complex and otherwise incomprehensible relationships can be clarified and structured. Thus, conclusions can be reached regarding large-scale connections that might otherwise appear irrelevant or excessively complicated. Moreover, meaningful information can be extracted by moving beyond individual perception when relatively little input data is available, because a network analysis does not necessarily focus on a specific location, but instead can encompass all types of relationships (Carrer-Neto, Hernández-Alcaraz, Valencia-García, & García-Sánchez, 2012).

Network analyses have also been used to improve the performances of recommender systems. Online social networking services can provide additional social relationship information about users (Kwon et al., 2009). Clustering in networks reduces the likelihood of information overload and improves the scalability of systems, because the similarity can only be calculated for users in the target cluster. Some studies have proposed clustering approaches based on social networks of users in order to derive CF recommendations (Colace, De Santo, Greco, Moscato, & Picariello, 2015; Seth et al., 2008), which can provide a rich source of information regarding user groups and their correlations. In addition, experimental evaluations have shown that using clustering techniques improves the prediction accuracy of user preferences (Pereira & Hruschka, 2015). Analyzing user activities in social network services (e.g., Facebook or LinkedIn) can provide e-commerce with the opportunity to create more personalized offers, and to help users cope with huge information overload problems (Zhou, Xu, Li, Josang, & Cox, 2012). In hybrid systems that combine CF and CBF methods, linear regression equations obtained from a social network graph are used to estimate the weight values of attributes (Debnath et al., 2008; Kushwaha, Mehrotra, Kalia, Kumar, & Vyas, 2016).