Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 2995–3013
Expert Systems with Applications
www.elsevier.com/locate/eswa

Collaborative recommender systems: Combining effectiveness and efficiency

Panagiotis Symeonidis *, Alexandros Nanopoulos, Apostolos N. Papadopoulos, Yannis Manolopoulos

Department of Informatics, Aristotle University, 54124 Thessaloniki, Greece

Abstract

Recommender systems base their operation on past user ratings over a collection of items, for instance, books, CDs, etc. Collaborative filtering (CF) is a successful recommendation technique that confronts the "information overload" problem. Memory-based algorithms recommend according to the preferences of nearest neighbors, and model-based algorithms recommend by first developing a model of user ratings. In this paper, we bring to surface factors that affect CF process in order to identify existing false beliefs. In terms of accuracy, by being able to view the "big picture", we propose new approaches that substantially improve the performance of CF algorithms. For instance, we obtain more than 40% increase in precision in comparison to widely-used CF algorithms. In terms of efficiency, we propose a model-based approach based on latent semantic indexing (LSI), that reduces execution times at least 50% than the classic CF algorithms.

Keywords: Recommender system; Collaborative filtering; Nearest neighbors

1. Introduction

The "information overload" problem affects our everyday experience while searching for knowledge on a topic. To overcome this problem, we often rely on suggestions from others who have more experience on the topic. However, in Web case where there are numerous suggestions, it is not easy to detect the trustworthy ones. Shifting from individual to collective suggestions, the process of recommendation becomes controllable. This is attained with the introduction of CF, which provides recommendations based on the suggestions of users who have similar preferences. Since CF is able to capture the particular preferences of a user, it has become one of the most popular methods in recommender systems.

Two types of CF algorithms have been proposed in the literature: memory-based algorithms, which recommend according to the preferences of nearest neighbors, and model-based algorithms, which recommend by first developing a model of user ratings. Both practical experience and related research have reported that memory-based algorithms (a.k.a. nearest-neighbor algorithms) present excellent performance, in terms of accuracy, for multi-value rating data. On the other hand, model-based algorithms are efficiently handle scalability to large data sets.

1.1. Motivation

Nearest-neighbor CF is influenced by several factors. Related research on CF, during the past decade, approached some of these factors. However, existing approaches may not be considered complete, because they examine the various factors only partially. More specifically, existing CF algorithms and their experimental evaluation focus only on parts of the CF process and do not

* Corresponding author.
E-mail addresses: symeon@csd.auth.gr (P. Symeonidis), ananopou@csd.auth.gr (A. Nanopoulos), papadopo@csd.auth.gr (A.N. Papadopoulos), manolopo@csd.auth.gr (Y. Manolopoulos).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.05.013

handle it as a whole. For the aspects that these partial considerations do not examine, they usually make choices, which our study demonstrates that can be misleading. Through our study we are also able to confirm that there exist dependencies between the factors affecting CF. Therefore, we have to perform an evaluation of the entire CF process in order to produce reliable conclusions.

Moreover, to handle scalability, we have to extend our findings for nearest-neighbor CF algorithms through a model-based approach. This approach will combine the effectiveness of the nearest-neighbor CF algorithms in terms of accuracy, with the efficiency in terms of execution time. Towards this direction, latent semantic indexing (LSI) is a technique that has been extensively used in informational retrieval. LSI detects latent relationships between documents and terms. In CF, LSI can be used to form users' trends from individual preferences, by detecting latent relationships between users and items. Therefore, with LSI, a higher level representation of the original user-item matrix is produced, which presents a three-fold advantage: (i) it contains the main trends of users' preferences, (ii) noise is removed, (iii) it is much more condensed than the original matrix, thus it favors scalability.

1.2. Contributions

In this work, first, we provide a thorough analysis of the factors involved in CF. Notably, we examine several similarity measures, various criteria for generating the recommendation list, the appropriateness of evaluation metrics, and the impact of CF in real-world applications, which is considered through user's satisfaction (measured with the popularity of recommended items) and the division of the ratings of the test user in past and future sets. During the analysis we identify choices that have been incorrectly adopted and new issues that have not been considered so far. As a result, we propose several extensions and new approaches, which substantially improve the entire CF process. Moreover, we propose a new model-based CF approach, which is based on LSI to produce a condensed model for the user-item matrix that handles the factor of scalability.

Our contributions are summarized as follows:

- The proposed approach examines the factors involved through the entire CF process. This helps to: (a) reveal fallacies in existing beliefs for several of them, (b) better analyze their impact and provide insights, and (c) synthesize a novel method, which substantially improve the effectiveness of CF in terms of accuracy.
- The previous findings are extended with an approach based on LSI. This way, execution times for CF are significantly reduced. Moreover, improvement is possible in effectiveness too, because the proposed model identifies the main trends and removes noise. Notice that differently from similar approaches in related work, we introduce the notion of pseudo-user in order to fold-in the vectors of target users in the produced model (related work incorrectly assumes the knowledge of target users during the building of the model).
- We carried out extensive experimental evaluation, which, to our knowledge, considers all the involved factors for the first time. Our experimental results demonstrate the superiority of the proposed methods (more than 40% improvements in terms of precision over widely-used CF algorithms and 50% in terms of execution times).

The rest of this paper is organized as follows. Section 2 summarizes the related work, whereas Section 3 contains the analysis of the CF factors. The proposed approaches are described in Sections 4 and 5. Experimental results are given in Section 6. Finally, Section 7 concludes this paper.

2. Related work

In 1992, the Tapestry system (Goldberg, Nichols, Brian, & Terry, 1992) introduced collaborative filtering (CF). In 1994, the GroupLens system (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994) implemented a CF algorithm based on common users preferences. Nowadays, it is known as user-based CF algorithm, because it employs users' similarities for the formation of the neighborhood of nearest users. Since then, many improvements of user-based algorithm have been suggested, e.g., (Breese, Heckerman, & Kadie, 1998; Herlocker, Konstan, Borchers, & Riedl, 1999).

In 2001, another CF algorithm was proposed. It is based on the items' similarities for a neighborhood generation of nearest items (Karypis, 2001; Sarwar, Karypis, Konstan, & Riedl, 2001) and is denoted as item-based CF algorithm.

Most recent work followed the two aforementioned directions (i.e., user-based and item-based). Herlocker, Konstan, and Riedl (2002) weight similarities by the number of common ratings between users/items. Deshpande and Karypis (2004) apply item-based CF algorithm combined with conditional-based probability similarity and cosine-similarity. Xue, Lin, and Yang (2005) suggest a hybrid integration of aforementioned algorithms (memory-based) with model-based algorithms.

All aforementioned algorithms are memory-based. Their efficiency is affected from scalability of data. This means that they face performance problems, when the volume of data is extremely big. To deal with this problem, many model-based algorithms have been developed (Breese et al., 1998). However, there are two conflicting challenges. If an algorithm spends less execution time, this should not worse its quality. The best result would be to improve quality with the minimum calculation effort.

Furnas, Deerwester, and Dumais (1988) proposed latent semantic indexing (LSI) in information retrieval area to deal with the aforementioned challenges. More specifically, LSI uses SVD to capture latent associations between the terms