Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 89 (2016) 450 – 456

Procedia Computer Science

Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)

A New Similarity Measure Based on Mean Measure of Divergence for Collaborative Filtering in Sparse Environment

Suryakant* and Tripti Mahara

IIT Roorkee, Roorkee 247 667, India

Abstract

Memory based algorithms, often referred to as similarity based Collaborative Filtering (CF) is one of the most popular and successful approaches to provide service recommendations. It provides automated and personalized suggestions to consumers to select variety of products. Typically, the core of similarity based CF which greatly affect the performance of recommendation system is to finding similar users to a target user. Conventional similarity measures like Cosine, Pearson correlation coefficient, Jaccard similarity suffer from accuracy problem under sparse environment. Hence in this paper, we propose a new similarity approach based on Mean Measure of Divergence that takes rating habits of a user into account. The quality of recommendation of proposed approach is analyzed on benchmark datasets: ML 100 K, ML-1 M and Each Movie for various sparsity levels. The results depict that the proposed similarity measure outperforms existing measures in terms of prediction accuracy.

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016

Keywords: Collaborative Filtering; Information Filtering; Recommendation System; Recommendation Accuracy; Similarity Measure.

1. Introduction

A Recommendation system (RS) is a necessity and a popular technology to handle information explosion. It serves as an information filtering tool and is commonly used to assist the target user to filter through a large pool of products and present only those products that are of user's interest. It gathers a large amount of data on the activities, inclinations, interest or taste of its client for a set of things i.e. movies, hardware item, garments and so forth and makes utilization of this gathered data to give suggestions to different clients.

In day to day life, we regularly depend on the opinion of like-minded individuals, individuals with comparable taste and preference or other trusted sources about the nature and quality of assets to get the suggestions for a product or item. A RS automates this word-of-mouth phenomenon and is generally utilized by the online shopping sites like Amazon to prescribe products of interest and by sites like Netflix to prescribe films to the users by giving personalization suggestions. The significance of RS increases inconceivably with the presence of long tail phenomenon (Anderson 2006). A physical retail store is characterized by shortage of resources. For instance, a fabric shop has constrained rack space and can show restricted assortment of items to a client. In the physical world it is impractical to

*Corresponding author. Tel: +918894969853; Fax: +91 1792 245362.
E-mail address: surya.dpt2015@iitr.ac.in

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016
doi:10.1016/j.procs.2016.06.099

customize a physical retail store for every individual client, as it is either represented by deals figures (most prevalent products) or it relies on upon master judgment. Despite what might be expected, an online retail store can make everything accessible to the client that exists. This peculiarity to makes things accessible between the offline and on-line world has been termed as the long tail[6] phenomenon. As per the long tail, offline retail stores just give the most popular things though on-line stores offer least popular items in addition most famous things. Henceforth[19], Park and Tuzhilin[19] proposed the need of a RS to prescribe things to individual clients online as it is unrealistic to present every accessible thing to the client or things that match the essence of clients.

Till date, Collaborative filtering (CF) is the most successful and widely employed approach in RS. Bobadilla *et al.*[3,9] is based on similar taste of users. It works on the fact that if a user had similar taste in past for a set of items, then they will share common taste in future. The information preferences for some resources are far more complex and hard to determine. It is due to the fact that sometimes preferences can't be just defined by using a set of keywords or by quality and taste. It can be obtained by observing the preference, behaviour or taste of other users. Tapestry[11] the first RS recommended documents, collected from newsgroup to a set of users by building database of contents and comments.

CF is one of the most widely used approaches to design a RS. However, sparsity is one of the major weaknesses of this prosperous approach. This problem inherently occurs in the system and is attributed to ever increasing number of users and items. Because of this numerous users may have evaluated or bought just couple of items from the total accessible items. Indeed, even extremely well known items may have been bought or evaluated by very few users. Hence it is difficult to compute similarity between users that leads to high sparsity in user-item ratings matrix. This affects the performance of a RS. In case, the system manages to evaluate similarity, their might exist a possibility that this similarity may be not reliable because of insufficient information processed. According to the density of matrix is lower than 1%.

In literature diverse similarity measures have been proposed but their performance is not very satisfactory for sparse matrix. Hence, the main aim of this paper is to build a new approach based on mean measure of divergence to find similarity between users to address the sparsity issue.

2. Related Work

One of the most critical factors that greatly affect the performance of CF is the computation of similarities between users. Cosine (COS), Pearson correlation coefficient (PCC), adjusted cosine measure (ACOS) and Spearman's rank correlation (SRC) are the generic traditional measures for similarity computation. These similarity measures defined in Table 1 Patra *et al.*[20].

Here the users are represented as vectors of objects based on their taste or preference history. The similarity between two users is characterized as the similarity of corresponding vectors. However literature has shown that the traditional measures do not properly utilize the user preferences (ratings data); especially when the available user item rating matrix is sparse or ratings data are not sufficient. Researchers have proposed some new similarity measures to improve the performance of a RS. Luo *et al.*[17] introduced local and global user similarity measure.

Bobadilla *et al.*[3] proposed a new similarity measure called JMSD by joining Jaccard *et al.*[14] and Mean squared-difference (MSD). Mean–Jaccard–Difference (MJD)[5] combination of six similarity measures. At that point, the neural network is utilized to tune the weights of each similarity measure. Choi & Suh[8] proposed another similarity measure that chooses neighbours dynamically for each different target item Pirasteh *et al.*[2] presented new weighted plans for customary similarity measures, which changed the symmetric similarity to asymmetric similarity. Jeong *et al.*[13] proposes to utilize an iterative message passing procedure for similarity updating. Gan and Jiang[10] utilize a power function to adjust user similarity scores.

The impact factor indicates about the preference of target user about an item. The penalty is imposed if ratings are on opposite side of the median. The popularity factor captures global information of the target item. The drawbacks of PIP[1] is addressed by Liu, Hu, Mian, Tian & Zhu, (2014) and motivated them to propose a new heuristic similarity model (NHSM). Patra *et al.*[20] proposed a similarity measure based on Bhattacharyya coefficient, it utilized all ratings given by a set of users. Chen *et al.*[7] propose new similarity measure using artificial immune network. In addition, a modified PCC formula is also proposed. To alleviated the sparsity problem of Collaborative recommender systems Leng *et al.*[15] proposed a novel similarity measure based on potential field.