Decision Support Systems 89 (2016) 17–27

Contents lists available at ScienceDirect

Decision Support Systems

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

A novel recommendation model with Google similarity

Tony Cheng-Kui Huang [a,*], Yen-Liang Chen [b], Min-Chun Chen [b]

[a] Department of Business Administration, National Chung Cheng University, 168, University Rd., Min-Hsiung, Chia-Yi, Taiwan, ROC
[b] Department of Information Management, National Central University, Chung-Li, Taiwan, ROC

ARTICLE INFO

Article history:
Received 6 October 2015
Received in revised form 2 May 2016
Accepted 5 June 2016
Available online 22 June 2016

Keywords:
Recommender system
Collaborative filtering
Normalized Google distance
Data mining

ABSTRACT

Previous studies on collaborative filtering have mainly adopted local resources as the basis for conducting analyses, and user rating matrices have been used to perform similarity analysis and prediction. Therefore, the efficiency and correctness of item-based collaborative filtering completely depend on the quantity and comprehensiveness of data collected in a rating matrix. However, data insufficiency leads to the sparsity problem. Additionally, cold-start is an inevitable problem concerning with how local resources are used as the basis for conducting analyses. This paper proposes a new idea by identifying an additional database to support item-based collaborative filtering. Regardless of whether a recommender system operates under a normal condition or applies a sparse matrix and introduces new items, this extra database can be used to accurately calculate item similarity. Moreover, prediction results acquired from two distinctive sets of data can be integrated to enhance the accuracy of the final prediction or successful recommendation.

1. Introduction

In the era of Web 2.0, users can exchange and share information through online media and open platforms including blogs, Facebook, and forums, thus generating a large amount of discussion and comment information in the World Wide Web (WWW). Such information can be collected and analyzed, enabling a recommender system access to additional information for outlining user profiles. Especially in the application of electronic commerce, the recommender system has used previous user purchase behavior and online review to aid users in quickly identifying suitable or interested items.

Currently, various recommender techniques have been developed [5,16,17]; hence, common recommender systems can be categorized as content-based and collaborative filtering systems. Content-based systems [3] mainly determine user preferences and recommend similar items according to items browsed or favored by the user. This type of system is disadvantageous for uncovering new customer preferences because only the customers' previous purchase behavior is accounted as the basis of recommendation. Therefore, scholars have proposed collaborative filtering [10,18,23], which is currently the most prevalent and discussed approach. Collaborative filtering defines clusters of users with similar preferences. Users in such clusters are defined as neighbors and exhibit common purchase or evaluation behavior that can be used to predict their future purchase behavior. The greatest difference between content-based and collaborative filtering systems is that the latter system derives user preferences through the purchase habit of a neighborhood, whereas conventional content-based systems only apply individual user data as the basis for recommending items.

In a collaborative filtering model, similarity calculation is the most crucial process and can be conducted to accurately determine similarity between items or customers, thus ensuring that subsequent predictions and inferences hold true. Collaborative filtering involves various methods that are mainly categorized as two types: user-based [11] and item-based [13,15, 21] collaborative filtering. The user-based method involves a similarity calculation approach to identify neighbor users with similar preferences or interests; hence, this method is named user-based or neighborhood-based collaborative filtering. However, the calculation time of this method increases with increasing number of users, thus affecting system efficiency [21]. Therefore, scholars have proposed item-based collaborative filtering according to the following basic assumption: An item that elicits user interest must be similar to another item that has received high user ratings. Overall, the item-based method calculates item–item similarity instead of user–user similarity.

Two problems have been frequently discussed regarding similarity calculation through item-based collaborative filtering:

(1) Sparsity Problem: When calculating the similarity between two items, a large amount of user information must be collected to ensure the reliability of the calculation results through increasing the sample number, thus preventing a scenario in which insufficient samples of user ratings yield unreliable calculation results. However, when information is insufficient, collecting a large amount of user information to avoid the sparsity problem is unviable in practice.

* Corresponding author. Tel.: +886 5 2729376; fax: +886 5 2720564.
E-mail address: bmahck@ccu.edu.tw (T.C.-K. Huang).

http://dx.doi.org/10.1016/j.dss.2016.06.005
0167-9236/

(2) Cold-Start: When incorporating a new item into the calculation process, the system cannot use the record of user ratings to determine the similarity of this item because no user has yet to rate it. Therefore, the system cannot estimate its user rating.

Previous studies have mainly used local resources as the basis for conducting analyses. For example, a rating matrix can be used to perform similarity and estimation analyses. Because only local resources are used to conduct analyses, the efficiency and correctness of a recommender process (e.g., the item-based collaborative filtering method) completely depend on the comprehensiveness of the rating matrix data. Lacking sufficient amount of local resources would lead to the sparsity problem. Additionally, cold-start is an inevitable problem concerning with how local resources are used as the basis for conducting analyses.

According to the discussion above, this paper proposes a new idea by using the WWW, a readily available large-scale database, to access global resources and calculate Google similarity, thus alleviating the problems concerning calculating similarity through local resources.

The concept of Google similarity involves using the large amount of discussion information posted on the WWW as well as the page count of the Google search engine to calculate similarity between items, thus improving the two problems of item-based collaborative filtering. More specifically, Google similarity is developed to measure a semantic similarity acquired from the number of hits returned by the Google search engine for a specified set of items. Items with the same or similar meanings tend to be close, while those with dissimilar meanings tend to be farther apart. We briefly describe the idea of Google similarity, and its formula will be introduced in the next section. First, we input two items separately to the search engine and obtain a maximum value with maximizing the two returned values. Second, we input both together and will have only one returned value. Ideally, the maximum value subtract the only one returned value should be almost equal to 0 if the two items are extremely close; otherwise, it should not be approach 0 if they are far.

According to the introduction above, we know that the sparsity problem occurs when local resources are insufficient for obtaining an authentic result regarding similarity; hence, this study applied Google similarity to adjust the calculated similarity, thereby increasing the usability of the calculation outcome. Furthermore, because conventional recommender systems only apply local resources and thus incur the cold-start problem, we also used Google similarity to recalculate similarly between items, thus resolving the problem of cold-start. Fig. 1 illustrates the overall system framework proposed in this study.

When estimating the user rating of an item, the system performs the following procedures:

(1) Use item-based collaborative filtering to calculate the rating of item I ($ac\_R_{x,non-rating}$).
(2) Use a Google similarity-based recommender system to calculate the rating of item I ($gs\_R_{x,non-rating}$).
(3) A parameter b is defined ($0 \le b \le 1$) to adjust the weights of $ac\_R_{x,non-rating}$ and $gs\_R_{x,non-rating}$ using the following formula: $(1-b)^* ac\_R_{x,non-rating} + b^* gs\_R_{x,non-rating}$. When $b = 1$, the system only uses the Google similarity-based collaborative filtering score as the estimated result. By contrast, when $b = 0$, only the item-based collaborative filtering score is used as the estimated result.

Finally, an experiment was conducted to verify the proposed system through integrating the Google similarity-based and item-based collaborative filtering methods. The results revealed that under four scenarios (the normal condition, sparsity problem, cold-start problem, and a simultaneous occurrence of the sparsity problem and the cold-start problem), the proposed system performed favorably.

As developing a new recommendation system, practitioners can consider adopting our proposed idea to strengthen the performance of recommendations. This kind of recommendation support system breaks through a thought that the data source of a support system not only is referring to the interior dataset but can be the exterior one. For customers, our proposed framework can support their purchase decisions with more appropriate recommendations.

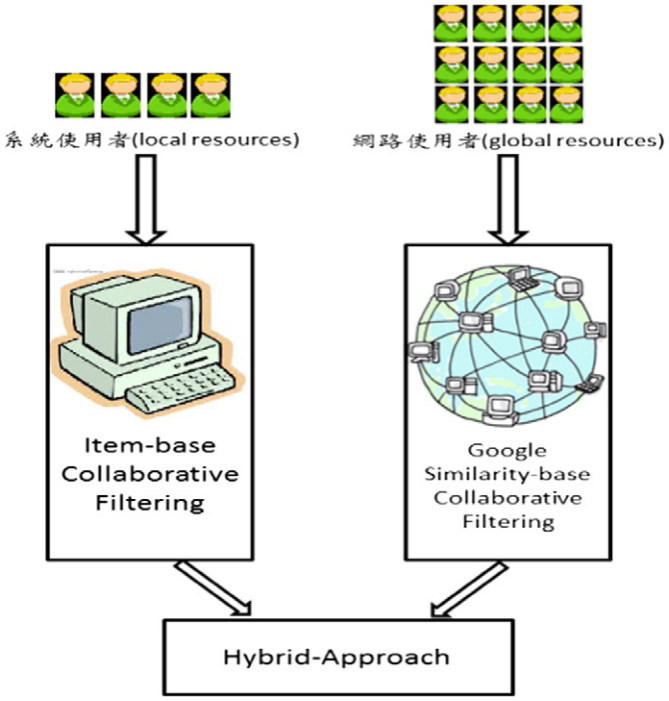

Fig. 1. System framework.

The remainder of this paper is organized as follows: we give a brief overview of the related works in Section 2. Section 3 introduces the research framework. In Section 4, we describe experiments using authentic datasets to evaluate the effectiveness of the proposed model. Conclusions are presented in Section 5.

2. Literature review

Section 2.1 first discusses the Normalized Goggle Distance (NGD). Section 2.2 elaborates the common methods applied by recommender systems such as content-based and collaborative filtering recommender systems.

2.1. Normalized Google distance

Google Search has become an irreplaceable tool for various users. When attempting to understand an unfamiliar keyword, it can be entered into the Google search engine to examine its definition by exploring other keywords displayed on a search page, enabling user to indirectly comprehend the definition of the keyword. For example, the term "rider" commonly appears together with the terms "horse" and "saddle". Consequently, the definition of "rider" can be inferred through its correlations with "horse" and "saddle". In other words, the Google search engine reveals that "rider" is associated with "horse" and "saddle".

In 2006, Cilibrasi and Vitanyi [6] used the Google search engine to investigate the correlation between two words or phrases. For example, when a computer attempts to learn the meaning of the word "hat", a database containing a vocabulary tree structure is constructed to represent correlations between words. Such a tree-structure can be established through two words. In addition, when searching for movie names by inputting the words "Captain America" and "The Avengers" into the Google search engine, a total of 17,000,000