Journal of Computational Science 26 (2018) 165–177

Contents lists available at ScienceDirect

Journal of Computational Science

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

Assigning reliability values to recommendations using matrix factorization

B. Zhu[a], F. Ortega[b], J. Bobadilla[c,*], A. Gutiérrez[c]

[a] Universidad Politécnica de Madrid, Spain & Beijing Institute of Technology, China
[b] U-tad Centro Universitario de Tecnología y Arte Digital, Madrid, Spain
[c] Universidad Politécnica de Madrid, Spain

ARTICLE INFO

Article history:
Received 25 June 2017
Received in revised form 5 April 2018
Accepted 10 April 2018
Available online 21 April 2018

Keywords:
Reliability
Confidence
Collaborative filtering
Matrix factorization

ABSTRACT

Providing a reliability value to each prediction and recommendation is very important in current recommender systems: Users should know which recommendations are reliable and which ones are risky. Despite its growing importance, research into collaborative filtering reliability has rarely been developed in the model-based area. This paper explains a matrix factorization-based architecture and method that provides a reliability value to each prediction/recommendation. The reliability values obtained have been put to the test, and, when applied, they show improvements in prediction and recommendation quality in different recommender systems; additionally, they provide a range of values that are understandable to users.

1. Introduction

Recommender Systems (RS) are playing an increasing part in technology to better the lives of every individual. RS popularity is growing fast, and users rely upon RS recommendations more and more every day; accordingly, RS scientists have the responsibility to enrich results by providing a reliability value to each prediction or recommendation. Users should know that some recommendations are more reliable than others; in fact, people like to know this information: Before we buy a product or book a hotel, beyond the averaged recommendation rating information (e.g. one to five stars), we carefully look at the number of opinions or votes other customers have given. We usually prefer a 4-star hotel based on 1000 opinions to a 5-star hotel based on 4 opinions.

In the above example, the number of opinions (or number of votes) plays the role of an easy to understand reliability value: Predictions and recommendation can be defined using the pair <averaged vote, number of votes>. Using Collaborative Filtering (CF) model-based RS, the prediction and recommendation processes are much more complex: Reliability values, as well as prediction values, must be obtained from hidden factors; fortunately, we can show recommendation results in the same simple pair representation: <prediction value, reliability value>. Reliability values will be mainly used to automatically (implicitly) filter the least reliable predictions or to show these reliability values to users and allow them to explicitly filter recommendations. Beyond this main use (beyond accuracy), model-based reliability values will open up some future opportunities such as: Improving cold-start results, explaining recommendations, extending the group recommendations methods, and merging sensor reliabilities with prediction reliabilities on the Internet of things RS.

This paper proposes a method to assign a reliability value to each prediction and, by extension, to each recommendation. This method is based on an architectural approach, rather than on a new mathematical model or algorithm. Moreover, the proposed method does not require any arbitrary parameters or specific algorithms; it is based on the use of any existing matrix factorization (MF) techniques. Therefore, it can be directly used in MF models, but not in other CF approaches, such as item or used-based cf. The method is based on the hypothesis that accuracy and reliability should be closely related: The more reliable a prediction the more accurate this prediction should be; similarly, experiments are based on this hypothesis to test the validity of the method. The tests' results show that reliable predictions are more accurate than non-reliable ones, and reliable recommendations present more precision.

In summary, we work on the hypothesis that the more suitable a reliability is, the better accuracy results will provide when applied: predictions with higher reliabilities should provide more accurate (less error) results, whereas we expect higher prediction errors on low reliability recommended items. This hypothesis follows the guidelines of the most relevant papers published in the area; by

* Corresponding author.
E-mail addresses: bo.zhu@alumnos.upm.es (B. Zhu), fernando.ortega@u-tad.com (F. Ortega), jesus.bobadilla@upm.es (J. Bobadilla), abraham@etsisi.upm.es (A. Gutiérrez).

https://doi.org/10.1016/j.jocs.2018.04.009
1877-7503/

way of example, in [41] it is stated: a) "Perhaps the most common measurement of confidence is the probability that the predicted value is indeed true", b) "We can design for each specific confidence type a score that measures how close the method confidence estimate is to the true error in prediction", and c) "Another application of confidence bounds is in filtering recommended items where the confidence in the predicted value is below some threshold".

Into paper [30], the reliability quality measure is tested by analyzing the way in which reliability values are related to prediction errors. In particular the "confidence curve", which is represented by the reliability values on the x-axis and the prediction errors on the y-axis, is evaluated. [19] is based on the same principle; it indicates: "This reliability measure is based on the usual notion that the more reliable a prediction, the less liable to be wrong". Finally, [18] states: "Evaluations of recommenders for this task must evaluate the success of high-confidence recommendations, and perhaps consider the opportunity costs of excessively low confidence".

This paper's most relevant contribution to RS state of the art is based on providing a measure of prediction and recommendation reliabilities, under the following restrictions: a) Not to rely on the use of additional data to the ratings cast: Social information, demographic data, temporal values, etc., and b) to use the most popular CF model: The MF method. Any reliability measure using data different to the rating matrix data will only be useful for those RS providing this specific type of data (social, geographical, etc.). The proposed method in this paper is valid for MF-based CF RS. Restriction b) our method (use of MF) allows the proposed solution to be applicable to most modern RS, which are model-based rather than memory-based and make widespread use of MF techniques.

Addressing the attainment of a reliability measure using MF techniques entails a difficulty that, as far as we know, has not been dealt with to date. In contrast to memory-based methods, where heuristic approaches have a place, in MF we are faced with a model-based method where the learning factors are hidden. Heuristic techniques do not seem to apply to factors in which the meaning is not known; a reasonable approach is the application of modern techniques of machine learning.

In this paper we propose an innovative architectural solution in the context of RS, with the same approach as some of the hierarchical architectural solutions currently used in the field of deep learning: We establish two differentiated levels of abstraction, both supported by MF as the machine learning method and arranged hierarchically. The first hierarchical level will provide us with the reliability associated with each existing rating prediction, while the second level will spread those values towards ratings not made.

As will be seen in the following section ("Related work"), as far as we know, there is no publication covering prediction and recommendation reliability measures on hidden factor-based models. Most of this field of research has focused on obtaining reliability measures using memory-based methods, in general, and datasets with social information or taxonomies, in particular. Section 3 explains the details of the proposed method; Section 4 sets out the experiments carried out to validate the quality of the results; Section 5 sets out the most relevant conclusions and proposes several future works. And finally, the related work references are provided.

2. Related work

The CF RS state of the art [39,5,45,10] includes a variety of research fields and application areas [11,43,29] where reliability has been relegated to a secondary level. The first papers in this area often used the term 'confidence', and they were located in the classic K-Nearest-Neighbours (*KNN*) method. This is a conceptually simple method, where recommendations are made by the set of *K* most similar (Nearest) users (Neighbours) to the active one. The idea behind the method is to recommend items that the active user does not know and that their neighbours have voted positively. Using *KNN*, confidence can be easily defined in terms of similarity of the neighbourhood to the active user or the number of ratings involved in each prediction.

When social networks emerged, RS data was enriched with tags, followers and followed information, as well as with all types of items and users' links. This was the starting point to create RS trust networks [46,36,9] obtaining users and items' reputation on the one hand, and prediction and recommendation trust measures on the other. Confidence/reliability values were obtained from trust and reputation values [40,27,28,14,33] and they were used to improve accuracy [2,19,29,26]. In RS where social information is available there is a possibility of using this additional information to obtain more reliable measures of confidence and trust. The next step in the reliability research path arrives when the Internet of things popularizes the existing context-aware RS [1,45,43]. Memory-based CF leading exponents are the *KNN* method and their necessary similarity measures [4,3], whereas the model-based CF leading exponents are the different matrix factorization methods [48,22]. Most of the reliability/confidence research has been carried out in the memory-based CF field [30,47,32], whereas the model-based CF has only occasionally covered prediction reliability [7].

A significant paper [19] in the CF reliability of predictions field provides a method to obtain specific reliability measures specially fitting the needs of different specific recommender systems. This method defines positive and negative factors: the greater the value of each positive factor for a prediction, the greater the value of the reliability of the prediction; the greater the value of each negative factor for a prediction, the lesser the value of the reliability of the prediction. This method mathematically combines positive and negative factors in order to compose a prediction's reliability measure. This paper also provides two specific KNN-based factors (one positive and one negative) to make a prediction reliability measure. This measure is tested, showing the expected behaviour: "the more reliable a prediction, the less liable to be wrong". We have taken this idea as baseline for our proposed method.

Authors of [32] provide a set of factors designed to obtain a reliability measure based on trust-aware information. First, they make the trust network of each active user, later the trust-based reliability measure made from [19] is used to evaluate the quality of the predicted rating, finally they reconstruct the trust networks for those of the users with lower reliability. Because the reliability measure proposed in [32] just works on datasets containing trust-aware information, it is not possible to use it as our baseline.

In order to test reliability measures' quality, [29] uses confidence curves: plots showing predictions' reliability values versus predictions' errors. Additionally, [29] provides a simple method to evaluate confidence curves. We have taken this concept: results into our paper provide this type of plots, as well as reliability/error correlation values. To test [29] quality measure, some simple methods have been put to the test: support and variability for user and item, resampling and injecting noise. Reliability estimation is especially valuable to weight imputation of predictions to the ratings matrix. To implement a novel imputation method, [38] provides a reliability measure. [38] uses its imputation method to estimate missing ratings and to impute them into the ratings matrix. The key idea is to differentiate between reliable and not reliable predictions: only reliable predictions are imputed into the ratings matrix.

Assigning uniform weights to the missing data is not an effective approach and it fails to keep up with the dynamic nature of online data [15]. Dynamically weighting missing data is analogous to establish a dynamic reliability measure (weight). The fast eALS (Alternating Least Squares) learning algorithm [15] is based on item popularity and it outperforms MF methods, using an incremental update strategy into the machine learning process. Item popular-