Enhanced user-user collaborative filtering recommendation algorithm based on semantic ratings

Wen Zhang*, Raja Chiky ** Manuel Pozo**

*Stanford University
zhangwen@cs.stanford.edu,
**ISEP Paris- LISITE Lab
firstname.lastname@isep.fr

1 Introduction

This paper presents a collaborative filtering recommendation algorithm that borrows ideas from content-based models by taking into account both the ratings and user preferences for item attributes. This is achieved by replacing each rating with its corresponding "semantic rating", which combines the original score and the user's historical preference level for the item's attributes (Pozo et al., 2016). This algorithm can achieve better results than a pure collaborative filtering counterpart oblivious to the intrinsic properties of each individual item.

We base our work on a previous experience (Pozo et al., 2016), which proposes the "semantic equation" as a transformation from an original rating to its corresponding semantic rating, thereby capturing a user's preference for an item's attributes. Since it is only a transformation of ratings, this technique can be used in different kinds of collaborative filtering algorithms. (Pozo et al., 2016) applied the semantic equation on top of a matrix factorization algorithm and found that the semantic approaches yielded better results in terms of precision, recall, and intra-list diversity. In this paper, we apply the same idea to a simple neighborhood method based on Pearson correlations. We validate (Pozo et al., 2016) and deeper shows the effect of these "semantic transfomation" by analyzing the "semantic amendments" only.

2 Semantic algorithms

The semantic algorithms that we propose are all based on the semantic equation, which turns a rating $r_{u,i}$ into its amended semantic rating $\mathfrak{r}_{u,i}$ (Pozo et al., 2016). For a user u and an item i, we define their "semantic amendment", which captures the user's historical preference for the item's attributes, as follows:

$$\Delta_{u,i} = \bar{r}_u \cdot \frac{\left|\sum_{a \in A(i)} C_{u,a} W_{t(a)}\right|}{|S(u)|}, \tag{1}$$

where $C_{u,a} = |\{i \in S(u) : a \in A(i)\}|$ denotes the number of times attribute a appears in all items rated by user u, and $W_{t(a)}$ denotes the "weight" of the attribute a's type. Note that $\Delta_{u,i}$

should be computed entirely from the training set, although it is not required that user u have rated item i.

The weight W_t of an attribute type t reflects its relevancy and is constant throughout the computation: an attribute type that is more relevant in making predictions should be assigned a higher weight. (Note that the weights belong to attribute *types* instead of individual attributes.). The weights are best chosen according to each attribute type's degree of relevancy obtained from PCA (Pozo et al., 2016).

The semantic equation can then be expressed as:

$$\mathfrak{r}_{u,i} = r_{u,i} + \Delta_{u,i}, \tag{2}$$

where $r_{u,i}$ can either be a rating originally present in the dataset, or be one predicted by an algorithm. Consequently, there is more than one place where we can apply Equation (2) in the training-testing process:

1. The *input-approach* semantic algorithm, results from applying Equation (2) to the calculation of Pearson correlations in the training stage.
2. The *output-approach* semantic algorithm results from applying Equation (2) to the prediction results: in the prediction stage, we transform each predicted rating $p_{u,i}$ into its semantic rating $\mathfrak{p}_{u,i} = p_{u,i} + \Delta_{u,i}$, which we use as the final prediction result.

3 CONCLUSION

We perform evaluations on the MovieLens-GroupLens dataset (Can, 2011), which consists of 2,113 users, 10,197 movies, and 855,598 ratings. It also contains six attributes: genres, directors, actors, countries, locations, and tags, with 112,881 distinct attribute values. We measure the precision, recall, and f-measure for a baseline algorithm (non semantic pearson algorithm) and the semantic algorithm. Our experiments demonstrate the effectiveness of using semantic ratings with a user-user collaborating filtering algorithm based on Pearson correlations. When information on items' attributes is present, making use of this information can often lead to better prediction performance by some metrics, even if we do not resort to a full hybrid algorithm.

The most evident strength of the semantic transformation is that it enables algorithms to suggest more items that are provably relevant: the more semantically-influenced an algorithm is, the higher precision and recall it yields. This finding is most directly supported by the fact that the pure output-approach semantic algorithm, which simply uses the semantic amendments as its predictions, gives substantially higher precision, recall, and f-measure.

References

(2011). *HetRec '11: Proceedings of the 2Nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, New York, NY, USA. ACM.

Pozo, M., R. Chiky, and E. Métais (2016). *Enhancing Collaborative Filtering Using Implicit Relations in Data*, pp. 125–146. Berlin, Heidelberg: Springer Berlin Heidelberg.