Evaluation of Attribute-aware Recommender System Algorithms on Data with Varying Characteristics

Karen H. L. Tso and Lars Schmidt-Thieme

Computer-based New Media Group (CGNM),
Department of Computer Science, University of Freiburg,
Georges-Koehler-Allee 51, 79110 Freiburg, Germany
{tso,lst}@informatik.uni-freiburg.de

Abstract. The growth of Internet commerce has provoked the use of Recommender Systems (RS). Adequate datasets of users and products have always been demanding to better evaluate RS algorithms. Yet, the amount of public data, especially data containing content information (attributes) is limited. In addition, the performance of RS is highly dependent on various characteristics of the datasets. Thus, few others have conducted studies on synthetically generated datasets to mimic the user-product relationship. Evaluating algorithms based on only one or two datasets is often not sufficient. A more thorough analysis can be conducted by applying systematic changes to data, which cannot be done with real data. However, synthetic datasets that include attributes are rarely investigated. In this paper, we review synthetic datasets applied in RS and present our synthetic data generation methodology that considers attributes. Furthermore, we conduct empirical evaluations on existing hybrid recommendation algorithms and other state-of-the-art algorithms using these variable synthetic data and observe their behavior as the characteristic of data varies. In addition, we also introduce the use of entropy to control the randomness of the generated data.

1 Introduction

Recommender systems use collaborative filtering to generate recommendations by predicting what users might be interested in, given some user's profile. Several prominent online commercial sites (e.g. amazon.com and ebay.com) offer this kind of recommendation services.

There are two different recommendation tasks typically considered: (i) predicting the ratings, i.e. how much a given user will like a particular item, and (ii) predicting the items, i.e. which N items a user will rate, buy or visit next (topN).

For RSs, nearest-neighbor methods, called collaborative filtering (CF ; [7]), is the prevalent method in practice. On the other hand, methods that rely only on attributes and disregard the rating information of other users, are commonly called the Content-Based Filtering (CBF). They have shown to perform very poorly. Yet, attributes usually contain valuable information; hence it makes it desirable to include attribute information in CF models – so called hybrid collaborative/content-based filtering methods.

There are many proposals on how to integrate attributes in CF for ratings. For instance, few others attempt linear combination of recommendation of CBF and CF predictions [5, 8, 10, 16]. There also exists methods that apply a CBF and a CF model sequentially, i.e. predict ratings by means of CBF and then re-estimate them from the completed rating matrix by means of CF [13]. There are also further proposals on how to integrate attributes when the problem is viewed as a classification problem [3, 4, 19]. As we lose the simplicity of CF, we do not consider those more complex methods here. We have selected three basic methods that predict items and try to keep the simplicity of CF, but still should improve prediction results.

When evaluating these recommendation algorithms, suitable datasets of users and items have always been demanding, especially when diversity of public data is limited. To compare the recommendation quality of different algorithms, it is not enough to evaluate the algorithms on just one or two datasets. Instead, one should investigate the behavior of the algorithms as systematic changes are applied to the data. Although there are already few attempts in generating synthetic data for the use in RS, to our best knowledge, there is no prior approach in generating synthetic data for evaluating recommender algorithms that incorporate attributes.

In this paper, we will make the following contributions: (i) we will propose our Synthetic Data Generator which produces user-item and user/item-attribute datasets and introduce the use of entropy to measure the randomness in the artificial data, (ii) we will survey some of the existing hybrid methods that consider attribute information in CF for predicting items. In addition, (iii) we will conduct empirical evaluations on three existing hybrid recommendation algorithms and other state-of-the-art algorithms using the generated synthetic data and observe their behavior when the characteristic of attribute data varies.

2 Related Works

One of the most widely known Synthetic Data Generators (SDG) in data mining is the one provided by the IBM Quest group [2]. It mimics the "real" world transactions in the retailing environment. It generates data with a structure and was originally intended for evaluating association rule algorithms. Later on, Deshpande and Karypis used this SDG for evaluating their item-based top-N recommendation algorithm [6]. Popescul *et.al* have proposed a simple approach by assigning a fixed number of users and items into clusters evenly and draw a uniform probability for each user and item in each cluster [17]. A similar attempt has been done for Usenet News [11, 14] as well as Aggarwal *et.al* for their horting approach [1]. Traupman and Wilensky tried to reproduce data by introducing skewed data to the synthetic data similar to a real dataset [20]. Another approach

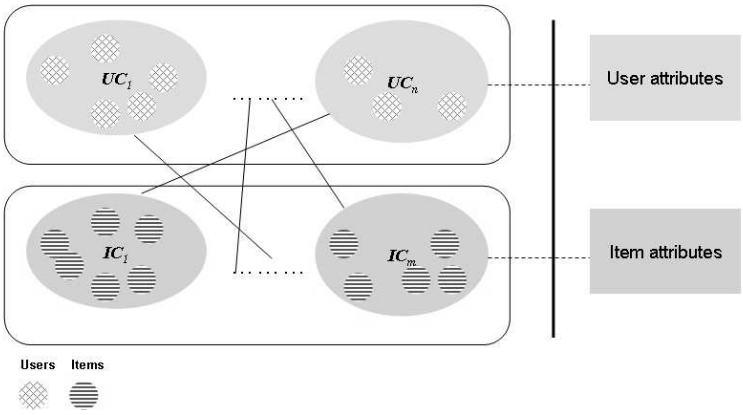

Fig. 1. Overview structure of synthetic data

is to produce datasets by first sampling a complete dataset and re-sample the data again by missing data effect [12].

The focus of this paper is to investigate SDG for CF algorithms which consider attributes. To the best of our knowledge, there is no prior attempt in examining SDGs for hybrid RS algorithms.

3 Synthetic Data Generator

The SDG can be divided into two phases: drawing distributions and sampling data. In the first phase, it draws distribution of User Cluster (UC) and Item Cluster (IC), next it affiliates UC or IC with user/item attribute respectively as well as to associate the UC and IC. Using these generated correlations, the users, items, ratings and item/user-attribute datasets can then be produced in the second phase. Fig. 1 presents an overview of how the artificial data are generally structured.

3.1 Drawing Distributions

To create the ratings and attributes datasets, we generate five random distributions models:

- $P(UC)$, how users are distributed in N number of UC.
- $P(IC)$, how items are distributed in M number of IC.
- $P(A|UC)\ \forall\ UC$, how user attributes (A) are distributed in UC.
- $P(B|IC)\ \forall\ IC$, how item attributes (B) are distributed in IC.
- $P(UC|IC)\ \forall\ IC$, how UC are distributed in IC.
- q be the probability that an item in IC_i is assigned to UC_j

The SDG first draws $P(UC)$ and $P(IC)$ from a Dirichlet distribution (with parameters set to 1). This asserts that the sum of $P(UC)$ or $P(IC)$ forms to one. $P(B|IC)$ shows the affiliation of item attributes with the item clusters by drawing from a special Chi-square distribution rejecting values greater than 1. Likewise, the correlation between UC and IC, $P(UC|IC)$, as well as the correlation between user attributes and user clusters, $P(A|UC)$, are done with similar manner. However, the attribute-aware CF algorithms we discuss in this paper do not take user-attributes into account. The overall drawing distributions process is summarized in (Algo. 1).

Algorithm 1 Drawing distribution

```
Input: $|A|, |B|, N, M, \epsilon_A, \epsilon_B, \epsilon_C$
Output: $P(UC), P(IC), P(A|UC), P(B|IC), P(UC|IC)$
  $h = 0$
  $P(UC) \sim Dir_{a_1,a_2...,a_N}$
  $P(IC) \sim Dir_{b_1,b_2...,b_M}$
repeat
    $P(B|IC)_h = S\chi^2 ED(|B|, M, h, \epsilon_B)$
    $P(UC|IC)_h = S\chi^2 ED(N, M, h, \epsilon_{IC})$
    $P(A|UC)_h = S\chi^2 ED(|A|, N, h, \epsilon_A)$
    $h = h + 0.1$
until $h < 1$
```

Algorithm 2 Drawing Special χ^2 distribution with specified entropy values

```
$S\chi^2 ED(n, m, H_{XY}, \epsilon_{XY})$ :
$d = 1$
repeat
  $P(X_i|Y_j) \sim \chi^2_d|_{[0,1]} \quad \forall i = 1...n, \forall j = 1...m$
  $d = d + 1$
until $|H(X|Y) - H_{XY}| < \epsilon_{XY}$
return $P(X|Y)$
```

By virtue of the randomness in those generated models, it is necessary to control or to measure the informativeness of these random data. Hence, we apply the Information Entropy and compute the average normalized entropy of the models.

$$H(X) = -\sum_{x \in \mathrm{dom}(X)} \frac{P(x) \log_2 P(x)}{\log_2 |\,\mathrm{dom}(X)|}. \tag{1}$$

The conditional entropy for the item-attribute data therefore is:

$$H(B_i|IC) = -\sum_{b=0}^{1} \sum_{j\in \mathrm{dom}\, IC} \frac{P(B_i = b, IC = j) \cdot \log_2 P(B_i = b|IC = j)}{\log_2 |\,\mathrm{dom}\, IC|} \tag{2}$$

In our experiment, $P(B|IC)$ is sampled eleven times for eleven different entropy values from 0 to 1 with 0.1 interval. By rejection sampling, $P(B \mid IC)$ is drawn iteratively with various Chi-square degrees of freedom until $H(B|IC)$ reaches desired entropies (Algo. 2). Other types of distribution have also been examined, yet, Chi-square distribution has shown to give the most diverse entropy range. We expect that as the entropy increases, which implies the data is less structured, the recommendation quality should decrease.

3.2 Sampling Data

Once these distributions have been drawn, users, items, ratings and item-attributes data are then sampled accordingly to those distributions. Firstly, users are assigned to user clusters by random sampling from $P(UC)$. Similar procedure, applies for sampling items. The user-item(ratings) data is generated by first sample $P(UC_l|IC_k)$ of users belonging to UC_l who prefer items in IC_k, then sample q portion of items of IC_k to these sampled users. The affiliation between items and attributes is done by sampling $P(B|IC)$ of items which contain attribute B. The same procedure can be applied to generate the user-attributes datasets. The overall sampling data process is summarized in (Algo. 3).

Algorithm 3 Sampling data

```
$uc_u \sim P(UC)$   user cluster of user u
$ic_i \sim P(IC)$   item cluster of item i
$oc_{l,k} \sim P(UC_l|IC_k)$   user of cluster $l$ who prefer item of cluster $k$
$o_{u,i} \sim binom(q)$   $\forall u, i : oc_{uc_u, ic_i} = 1$   occurrence of user of $uc_u$ prefers item of $ic_i$
$o_{u,i} = 0$   else
$b_{i,t} \sim P(B_t|IC = ic_i)$   item $i$ contains attribute $t$
```

4 Hybrid Attribute-Aware CF Methods

Here, we discuss three existing hybrid methods [21], which will be evaluated using the data generated from the SDG.

1. Sequential CBF and CF (Adapted Content-Boosted CF),
2. Joint Weighting of CF and CBF, and

3. Attribute-Aware Item-Based CF.

Sequential CBF and CF is the adapted version of an existing hybrid approach, Content-Boosted CF, originally proposed by [13] for predicting ratings. This method has been conformed to the predicting items problem here. It first uses CBF to predict ratings for unrated items and then filters out ratings with lower scores (i.e. keeping ratings above 4 on a 5-point scale) and applies CF to recommend topN items.

Joint Weighting of CF and CBF (Joint-Weighting CF-CBF), first applies CBF on attribute-dependent data to infer the fondness of users for attributes. In parallel, user-based CF is used to predict topN items with ratings-dependent data. Both predictions are joint by computing their geometric mean.

Attribute-Aware Item-Based CF (Attr-Item-based CF) extends item-based CF [6]. It exploits the content/attribute information by computing the similarities between items using attributes thereupon combining it with the similarities between items using ratings-dependent data.

All three approaches recommend items that contain the highest frequency of their neighboring items. For the last two algorithms, λ is used as a weighting factor to vary the significance applied to CF or CBF.

5 Evaluation and Experimental Results

In this section, we present the evaluation of the selected attributes-aware CF algorithms using artificial data generated by SDG discussed in Section 3 and compare their performances with their corresponding non-hybrid base models, which do not integrate attributes, i.e. user-based and item-based CF, as well as to observe the behavior of the algorithms after supplement of attributes.

Metrics Our paper focuses on the item prediction problem, which is to predict a fixed number of top recommendations and not the ratings. Suitable evaluation metrics are Precision, Recall and F1.

Similar to Sarwar *et al.* [18], our evaluations consider any item in the recommendation set that matches any item in the testing set as a “hit”. F1 measure is then used to combine Precision and Recall into a single metric.

$$\begin{aligned}\text{Precision} &= \frac{\text{Number of hits}}{\text{Number of recommendations}}\\ \text{Recall} &= \frac{\text{Number of hits}}{\text{Number of items in test set}}\\ \text{F1} &= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}\end{aligned}$$

Parameters Due to the nature of collaborative filtering, the size of neighborhood has significant impact on the recommendation quality [9]. Thus, each of the randomly generated data should have an assorted neighborhood sizes for each method. In our experiments, we have selected optimal neighborhood sizes and λ parameters for the hybrid methods by means of a grid search. See Table 1. Lambda is used to weight the contribution of attribute-dependent and rating-dependent models. Threshold and max, for the Sequential CBF-CF are set to 50 and 2 accordingly as chosen in the original model [13]. For more detail explanation of the parameters used in those algorithms, please refer to [21] and [13].

Table 1. The parameters chosen for the respective algorithms.

Method	Neighborhood Size	λ
user-based CF	35-50	–
item-based CF	40-60	–
joint weighting CF–CBF	35-50	0.15
attr-aware item-based CF	40-60	0.15

As our algorithms do not consider user attributes, our SDG only generates models for item attributes. The parameter settings for our experiments are summarized in Table 2.

Table 2. The parameters settings for the synthetic data generator

Description	Symbol	Value
Number of users	n	250
Number of items	m	500
Number of User Clusters	N	5
Number of Item Clusters	M	10
Number of Item Attributes	$\|B\|$	50
Probability of i in IC assigned to a UC	q	0.2

Experimental Results In our experiments, we have generated five different trials. For each trial, we produce one dataset of user-item (ratings) and eleven different item-attributes datasets with increasing entropy from 0-1 with 0.1 interval, by rejection sampling. In addition, to reduce the complexity of the experiment, it is assumed that the correlation between the user and item clusters to be fairly well-structure and have a constant entropy of 0.05. The results of the average of five random trials where only item-attributes with entropy of 0.05 are presented in Fig. 2.

As shown in Fig. 2, Joint-Weighting CF-CBF achieves the highest Recall value by around 4% difference w.r.t. its base method. On the other hand, Attr-

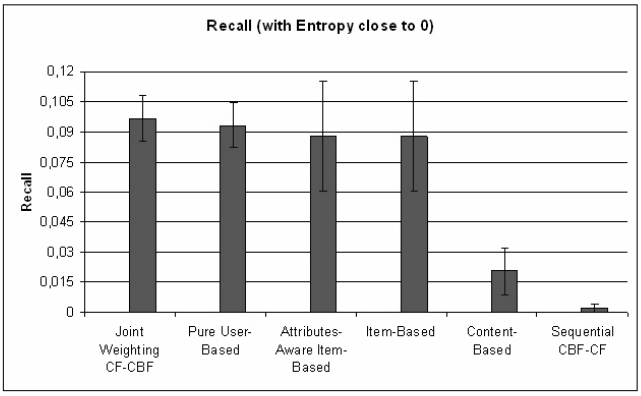

Fig. 2. Recall by selecting item-attributes with entropy ≤ 0.05

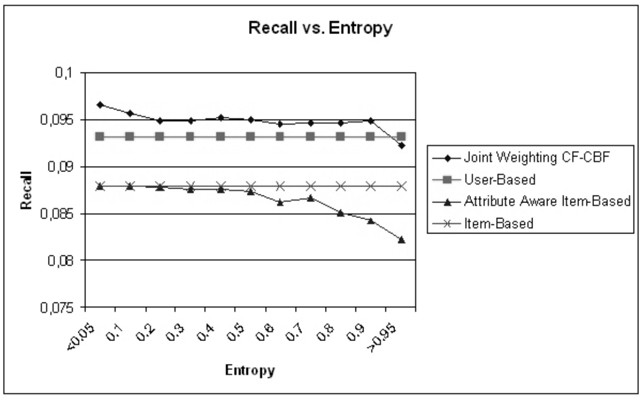

Fig. 3. Recall vs. Entropy ranging from 0-1

Item-based CF does not seem to be effective at all as attributes are appended to its base model. It also has a very high standard deivation. This suggests that the algorithms to be rather unstable and unreliable. Although Melville *et al.* [13] reported that Content-Boosted CF performed better than user-based and pure CBF for ratings, it fails to provide quality top-N recommendations for items in our experiments. Therefore, we will focus our evaluation on the other two algorithms in the rest of the paper.

As the aim of the paper is to examine the behavior of the models as the characteristic of data varies, what is more important is to observe the performance as entropy varies. As anticipated, the recommendation quality increases, when there exists more structure in the data. The results of an average of five random trials of item-attribute datasets with eleven various entropies are presented in Fig. 3.

We can see that for both Attr-Item-based CF and Joint-Weighting CF-CBF algorithms, the quality of recommendation reaches its peaks when the entropy approaches zero and it gradually decreases as entropy increases. As for Attr-Item-based CF, although it carries the right entropy trend, its peak does not surpass its base model and the quality drops gradually below its base model, which does not make use of attributes. On the other hand, for Joint-Weighting CF-CBF, the value of recall descends gradually as the entropy raises, still the recall maintain above its base-model until entropy approaches 1 where recall plummets to below its base-line score.

6 Conclusions and Future Works

The aim of this paper is to conduct an empirical evaluation on three existing hybrid recommendation models and other state-of-the-art algorithms with data generated by the SDG presented in this paper. All algorithms discussed here focus on the predicting items problem. Joint-Weighting CF-CBF, appears to enhance recommendations quality when reasonable amount of informative attributes are presented. The other algorithms do not seem to be sensitive to attributes. Yet, we expect the outcomes could be ameliorated by adding more structural dependency between clusters. In addition, currently the data are only controlled by the entropy of item-attribute datasets; however, other distributions such as the user-item data should also be investigated when various entropies are considered. Furthermore, more extensive experiments should be done to examine the effect of varying other parameters settings and to conduct an empirical evaluation with models that predict ratings.

References

1. Aggarwal, C. C., Wolf, J. L., Wu, K.-L. and Yu, P. S.: Horting hatches an egg: A new graph-theoretic approach to collaborative filtering. In Proceedings of ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York (1999)
2. Agrawl, R. and Srikant, R.: Fast algorithms for mining association rules. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB). Morgan Kaufmann (1994) 487-499
3. Basilico, J. and Hofmann, T.: Unifying collaborative and content-based filtering. In Proceedings of the 21st International Conference on Machine Learning. Banff, Canada (2004)
4. Basu, C., Hirsh H., and Cohen, W.: Recommendation as classification: Using social and content-based information in recommendation. In Proceedings of the 1998 Workshop on Recommender Systems. AAAI Press, Reston, Va. 11-15 (1998)
5. Claypool, M., Gokhale, A. and Miranda T.: Combining content-based and collaborative filters in an online newspaper. In Proceedings of the SIGIR-99 Workshop on Recommender Systems: Algorithms and Evaluation (1999)
6. Deshpande, M. and Karypis, G.: Item-based top-N recommendation algorithms, ACM Transactions on Information Systems **22/1** (2004) 143–177

7. Goldberg, D., Nichols, D., Oki, B. M. and Terry, D.: Using collaborative filtering to weave an information tapestry. Commun. ACM **35** (1992) 61–70
8. Good, N., Schafer, J.B., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J., and Riedl, J.: Combining Collaborative Filtering with Personal Agents for Better Recommendations. In Proceedings of the 1999 Conference of the American Association of Artificial Intelligence (AAAI) (1999) 439-446
9. Herlocker, J., Konstan, J., Borchers, A., and Riedl, J.: An Algorithmic Framework for Performing Collaborative Filtering. In Proceedings of ACM SIGIR'99. ACM press (1999)
10. Li, Q. and Kim, M.: An Approach for Combining Content-based and Collaborative Filters. In Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages (ACL) (2003) 17–24
11. Konstan, J. A., Miller, B. N. , Maltz D., Herlocker, J. L., Gordon, L. R. and Riedl, J.: Group-Lens: Applying collaborative filtering to usenet news. Commun. ACM 40 (1997) 77-87
12. Marlin, B., Roweis, S. and Zemel, R.: Unsupervised Learning with Non-ignorable Missing Data. In Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS) (2005) 222–229
13. Melville, P., Mooney, R. J. and Nagarajan, R.: Content-Boosted Collaborative Filtering for Improved Recommendations. In Proceedings of the Eighth National Conference on Artificial Intelligence(AAAI-2002). Edmonton, Canada (2002) 187–192
14. Miller, B. N., Riedl, J. and Konstan, J. A.: Experiences with GroupLens: Making Usenet useful again. In Proceedings of the 1997 USENIX Technical Conference (1997)
15. MovieLens: Available at http://www.grouplens.org/data (2003)
16. Pazzani, M. J.: A framework for collaborative, content-based and demographic filtering. Artificial Intelligence Review 13(5-6):393408 (1999)
17. Popescul, A., L.H. Ungar, D.M. Pennock, and S. Lawrence: Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (2001) 437–444
18. Sarwar, B. M., Karypis, G., Konstan, J. A. and Riedl, J.: Analysis of recommendation algorithms for E-commerce. In Proceedings of the 2nd ACM Conference on Electronic Commerce (EC). ACM, New York (2000) 285–295
19. Schmidt-Thieme, L.: Compound Classification Models for Recommender Systems. In Proceedings of the IEEE International Conference on Data Mining (ICDM). New Orleans, USA (2005) 559–570
20. Traupman, J. and Wilensky, R.: Collaborative Quality Filtering: Establishing Consensus or Recovering Ground Truth?. In Proceedings of WebKDD 2004: KDD Workshop on Web Mining and Web Usage Analysis, in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), August 22-25 2004. Seattle, WA (2004)
21. Tso, H. L. K., Schmidt-Thieme, L.: Attribute-Aware Collaborative Filtering. In Proceedings of the 29th Annual Conference of the German Classification Society 2005. Magdeburg, Germany (2005)