Time Dependency of Data Quality for Collaborative Filtering Algorithms

Toon De Pessemier   Simon Dooms

Tom Deryckere   Luc Martens

WiCa group, Dept. of Information Technology, IBBT-Ghent University
Gaston Crommenlaan 8 box 201, B-9050 Ghent, Belgium
{Toon.DePessemier, Simon.Dooms, Tom.Deryckere, Luc.Martens}@intec.ugent.be

ABSTRACT

The efficiency of personal suggestions generated by collaborative filtering techniques is highly dependent on the quality and quantity of the available consumption data. Extending data sets with additional consumption data (from the past) might enrich the user profiles and generally leads to more accurate recommendations. Although if a considerable amount of profile information is already available and detailed personal preferences can be derived, supplementary consumption data may not have any (or a very limited) added value for the recommendation algorithm. These additional consumption data increase the required storage capacity and the computational load to generate the personal recommendations. Moreover, since personal preferences and the relevance of content items may vary over time, older consumption data might be outdated and lead to inaccurate recommendations. Therefore, we investigate which consumption data are (the most) relevant to feed the conventional collaborative filtering algorithms. For provider-generated content systems, we demonstrate that the accuracy of collaborative filtering algorithms increases by extending user profiles with additional older consumption data. In contrast, we witness the opposite effect for user-generated content systems: involving older consumption data has a negative influence on the recommender accuracy. These results are important for website owners who intend to employ a recommendation system at a minimum storage and computation cost.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Performance

Keywords

Recommender systems, Collaborative filtering, Data quality

1. INTRODUCTION

Despite the popularity of collaborative filtering (CF) techniques, the required storage and computational cost might seriously increase for large-scale online systems due to the quadratic nature of the neighbourhood selection process [7]. Therefore, we opted for a different approach by investigating the minimal amount of consumption data required as input to acquire optimal recommendations. If as soon as a certain amount of consumption data is obtained, additional consumption data do not contribute to a further increase in recommendation accuracy, then these additional data can be considered as useless for a recommender. As a result, these additional consumption data do not have to be stored anymore and can be excluded from the calculations, which reduces the storage and calculation costs for website owners. The remainder of this paper is organised as follows: Section 2 presents the applied research process and describes in detail the employed data sets, evaluation method and evaluation metrics. In Section 3, we discuss the results obtained by three different experiments on two representative data sets. Finally, we offer a brief conclusion and point out interesting future work in Section 4.

2. METHODOLOGY

To determine the minimal amount of consumption data that is required to acquire optimal recommendations, we calculated recommendations based on different versions of the consumption history in successive iterations. These versions vary in the number of consumptions and as a result in the sparsity of the data matrix, which can be derived from the consumption history. The accuracy of the generated recommendations will be compared for the successive iterations based on evaluation metrics which are generated by an offline analysis on a test set of consumptions.

2.1 Data Set Evaluation

To estimate the effectiveness of personal recommendations, two different evaluation methods are possible. On the one hand, online evaluations measure how the user interacts (e.g. clicking, buying behaviour) with the presented recommendations on a running service. Offline evaluations, on the other hand, partition historical consumption data in a training set and a test set. Based on the training set, the recommender has to predict the consumption behaviour in the test set. Although online evaluation methods are closest to reality, we opted for an offline evaluation based on data sets because such an evaluation is fast, reproducible and com-

monly used in recommendation research. We performed the evaluations with two different data sets: one data set that contains consumption behaviour of provider-generated content and one with consumption behaviour of user-generated content. To evaluate the recommendations for provider-generated content, we opted for a data set that is commonly used to bench-mark recommendation algorithms, namely the (medium-size) Movielens data set [8]. This set contains 3,706 items (videos), 6,040 active users and 1,000,209 distinct consumptions, logged from April 2000 until February 2003, which we ordered chronologically and transformed to a binary scale.

Although user-generated content systems are increasingly popular on the Internet, data sets with consumption behaviour of user-generated content are less common in recommendation research. Therefore, we used a less well known data set with user-generated content behaviour collected at PianoFiles [1]. PianoFiles is a user-generated content site that offers users the opportunity to exchange, browse and search for sheet music they like to play. Currently, users can manage their personal collection of sheet music on PianoFiles but they do not yet receive personal recommendations. The main consumption behaviour, used to feed the recommendation algorithm, consists of the personal collections of the users. Each addition to a personal collection is used to populate the consumption matrix. This full data set contains 110,809 items (sheets), 85,537 active users and 1,211,572 distinct consumptions in chronological order, logged from May 2003 until September 2009. So, the consumption behaviour is gathered over a longer period of time on PianoFiles than on Movielens. Nevertheless, the time window of these two data sets is comparable since most consumption behaviour of PianoFiles was registered during the last 3 years. Moreover, Movielens and PianoFiles contain approximately the same number of consumptions in their data set. However, due to the typical content (production) characteristics of user-generated content systems [3], the sparsity of the PianoFiles data set is much higher than that of Movielens, which makes it more difficult to calculate accurate predictions for PianoFiles. For instance, approximately 50% of the PianoFiles users are suffering from the cold start problem, having a limited profile of less than 5 sheets in their collection. Conversely, a considerable part of the PianoFiles users (approximately 14%) has a collection containing more than 20 sheets, which enables a recommendation system to calculate detailed suggestions.

2.2 Evaluation Method & Metrics

For evaluation purposes, we used the 50% most recent consumption data as the test set. The remaining 50% of the consumption records were used as input data. This offline evaluation methodology, in which a data set is chronologically split in training set and test set, is commonly used to evaluate recommendation algorithms [4]. In order to study the performance of the algorithm with data of different sparsity levels, we composed ten different training sets by eliminating 0%, 10%, 20%, ... up to 90% of the input data. Afterwards, the recommendation algorithm used these different training sets in successive iterations to generate personal suggestions. As is commonly done for the evaluation of recommendations under sparse data [6], the test set was first filtered to include only consumptions that are predictable with the input data as a priori knowledge. A consumption of an item that is
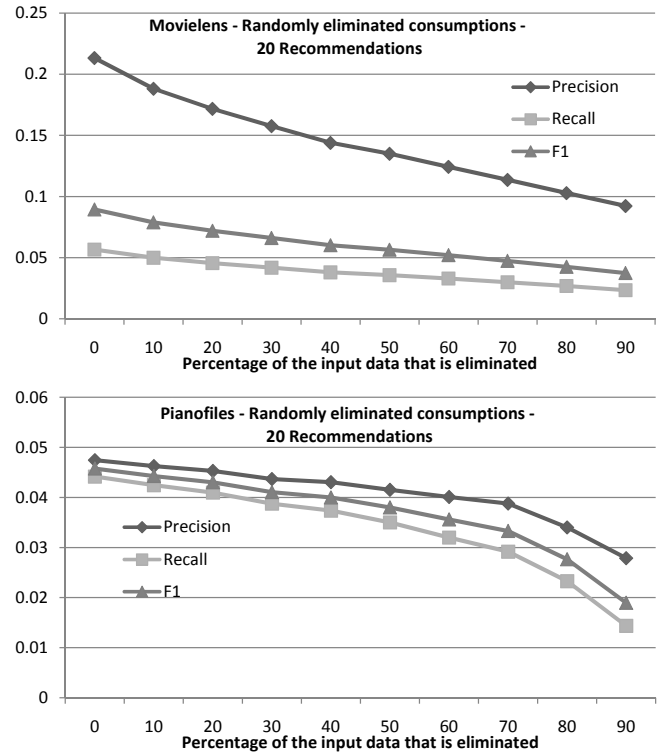

Figure 1: The evaluation of the UBCF with randomly depleted input data as training set. The graphs visualize the evolution of the precision, recall and F1 as a function of the percentage of input data that is eliminated.

not contained in the input data or a consumption of a user without any consumption behaviour in the input data is impossible to predict with CF techniques. All users of this filtered test set were included into a set of target consumers. For each of these consumers, the algorithm generated five ordered lists of 10, 20, 30, 40 and 50 recommendations respectively, which were compared with the test set. (Only the results for 20 recommendations per user are shown in this paper, since the other results illustrate the same conclusions.) One of the most frequently used error metrics is the Root Mean Squared Error (RMSE) [5, 2], which was also adopted by the official Netflix contest. However, the Netflix contest was mainly focused on predicting accurate ratings for an entire set of items, while web-based applications are mostly interested in providing the users with a short recommendation list of interesting items [2]. To evaluate this top-N recommendation task, i.e. a context in which we are not interested in predicting user ratings precisely, but rather in giving an ordered list of N attractive items to the users, error metrics are not meaningful. Therefore, information-retrieval classification metrics, which evaluate the quality of a short list of recommendations, are most suitable. The most popular classification accuracy metrics are precision, recall and F1 [2], which we adopted in this research.

3. EXPERIMENTAL RESULTS

Compared to an item-based CF algorithm, a user-based strategy achieves much better results for our sparse data set,

obtained from PianoFiles. A possible explanation for this is the high number of items (compared to the number of users), which makes it difficult to find optimal neighbourhoods of similar items. Therefore, we applied the standard user-based CF algorithm (UBCF) with the cosine similarity as a measure to compare user profile vectors, for all our bench-marks [9]. Because of the difference in the sparsity level between the provider-generated and the user-generated content system, the absolute values of the evaluation metrics can not be compared. However, the evolution of these metrics under different sparsity levels is an interesting characteristic to study the time dependency.

3.1 Randomly Removing Consumptions

In a first experiment, personal suggestions were generated with a UBCF algorithm based on randomly depleted input data. In a primary iteration, the initial input data without modification were used as a training set to obtain a baseline situation. Next, the input data were depleted by randomly eliminating 10%, 20%, until 90% of the consumptions, in successive iterations.

Figure 1 shows that the evaluation metrics decrease linearly for the Movielens data set as more consumptions are removed from the input data. Indeed, the reduced amount of available consumption information and the accordingly confined user profiles lead to more inaccurate recommendations, which lower the precision and the recall value. Moreover, the recall value might also decrease if the CF algorithm is no longer able to generate (sufficient) suggestions for users with a severely depleted profile. As with Movielens, the evaluation metrics obtained with the PianoFiles data set decrease linearly for the first iterations. In contrast, the evaluation metrics drop suddenly after a certain amount of consumption data was removed. This can be explained by the sparsity of the data set for user-generated content systems: as soon as the amount of available consumption data is below a certain threshold, the data matrix is too sparse to produce accurate recommendations.

3.2 Chronologically Removing Consumptions: All Users

In a second experiment, the input data were depleted chronologically. Again, a primary iteration, which was based on all the consumptions of the input data, represented the baseline situation. In the following iterations, 10%, 20%, . . . up to 90% of the consumptions were eliminated from the input data, starting with the oldest consumptions. So, in the second iteration for example, the most recent 90% of the input data were used as training information for the recommender.

Figure 2 shows that the evaluation metrics decrease linear for the Movielens data set just like the random depletion process of the first experiment. However, the comparison of Figure 1 and 2 indicates this deterioration is less severe if the oldest data are removed instead of random data. This proves that the oldest consumption behaviour has the lowest information value for recommendation systems. The results that were obtained with the PianoFiles data set deviate from the previously observed behaviour. Figure 2 illustrates that if older consumption data are removed from the input, the recall value decreases, but unlike the previous experiment, the precision improves. This increasing precision value can be attributed to two influence sources. Firstly, since the UBCF
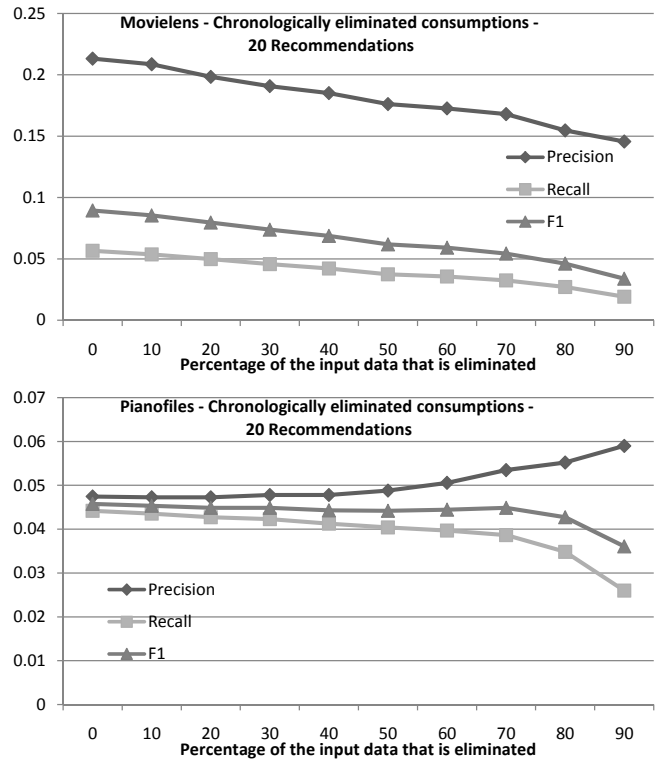

Figure 2: The evaluation of the UBCF with chronologically depleted input data as training set. The graphs visualize the evolution of the precision, recall and F1 as a function of the percentage of input data that is eliminated.

algorithm can not generate recommendations for users without consumption behaviour in the training data, the generated recommendations are limited to the users that were active in the most recent time period for which consumption behaviour is available. This way, the recommender has to focus on generating recommendations for recently-active users, whose behaviour might be easier to predict than that of users without recent activity on the website. So, in successive iterations, the number of users for whom recommendations are generated is reducing, which might increase the precision of these recommendations. Secondly, eliminating old consumption data from the input might clear the user profiles by reducing the number of obsolete items. User-generated content can be very transient, evolving from very popular to totally outdated in a very short period. Removing old consumptions from the input reduces the risk that obsolete items will be recommended, which might increase the precision of the recommendations as well.

3.3 Chronologically Removing Consumptions: Recently-Active Users Only

To distinguish the two previously described influences, which increase the precision value for the PianoFiles data set in successive iterations, we modified the previous experiment. In this experiment, the recommender had to produce recommendations only for users that were active in the most recent 10% of the input data, instead of generating recommendations for all the users with an available profile. For
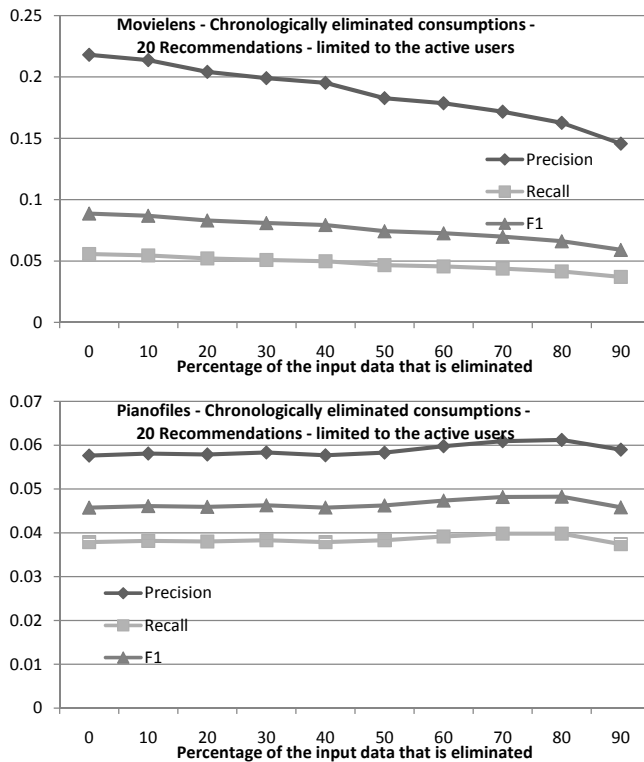
Figure 3: The evaluation of the UBCF, limited to the users that were recently active, with chronologically depleted input data as training set. The graphs visualize the evolution of the precision, recall and F1 as a function of the percentage of input data that is eliminated.

this set of users, it is still possible to generate a personal profile if the oldest 10%, 20% up to 90% of the consumption data are removed from the input. So, the number of users for whom the UBCF algorithm can generate recommendations remains unchanged for successive iterations. This way, the first influence source, which was described in the previous section, is eliminated in this experiment. So, Figure 3 merely reveals the effect of removing old consumption data from existing user profiles on the accuracy of the recommendations. The graph confirms that older consumption data still have an important information value for the Movielens data set, since the precision and recall decrease if these data are removed from the input. However, we witness the opposite effect for the PianoFiles data set: precision and recall remain stable as older consumption data are removed in successive iterations. We even notice a slight increase for the last iterations (70%, 80% and 90% removed data), which means that the accuracy of the recommendations improves if old information is not considered in the algorithm.

4. CONCLUSION & FUTURE WORK

In this research, we investigated the necessity to utilize all the available consumption data to generate recommendations with a traditional collaborative filtering algorithm. Our results showed that additional consumption data from the past can provide a positive contribution to the accuracy of the recommendations for provider-generated con-

tent (Movielens data set). For user-generated content (PianoFiles data set) however, the results showed that older consumption data have no added information value for traditional collaborative filtering algorithms. Inserting older consumption data in user profiles of user-generated content systems might even lead to less accurate recommendations. We consider the transient popularity of user-generated content responsible for this decline. This conclusion has important consequences for recommendation systems for user-generated content. Since older consumption data have no added value for the recommender, there is no need to store these data or incorporate the data into the calculations. This reduces the storage and computational burden for large user-generated content systems which offer personal suggestions to their customers. In future research, we will verify our conclusion with other data sets and an online evaluation. Moreover, we will try to generalize the conclusions for more types of recommendation algorithms. Finally, we will investigate the optimal time period to log the consumption data for recommendation purposes.

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] T. Bonte. PianoFiles / Free sheet music for piano, 2003. Available at http://www.pianofiles.com/.

[2] E. Campochiaro, R. Casatta, P. Cremonesi, and R. Turrin. Do metrics make recommender algorithms? *Advanced Information Networking and Applications Workshops, International Conference on*, 0:648–653, 2009.

[3] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 1–14, New York, NY, USA, 2007. ACM.

[4] C. Hayes, P. Massa, P. Avesani, and P. Cunningham. An on-line evaluation framework for recommender systems. In *Workshop on Personalization and Recommendation in E-Commerce (Malaga)*. Springer Verlag, 2002.

[5] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.

[6] Z. Huang, D. Zeng, and H.Chen. A link analysis approach to recommendation with sparse data. In *AMCIS 2004: Americas Conference on Information Systems*, New York, NY, USA, 2004.

[7] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, 2003.

[8] M. Harper et al. Grouplens research - movielens data sets, 2006. Available at http://www.grouplens.org.

[9] T. Segaran. *Programming collective intelligence*. O'Reilly, 2007.

Search | SEARCH

ACM Recommender Systems 2010

We are pleased to invite you to participate in this premier annual event on research and applications of recommender technologies.

The 4th ACM Conference on Recommender Systems builds on the success of Recommenders 06 Summer School in Bilbao, Spain; the 1st Conference in Minneapolis, USA; the 2nd Conference in Lausanne, Switzerland; and the 3rd Conference in New York, USA. In these events many members of the practitioner and research communities valued the rich exchange of ideas made possible by the shared plenary sessions. The 4th conference will promote the same close interaction among practitioners and researchers, reaching a wider range of participants including those from Europe and Asia. Published papers will go through a full peer review process. The conference proceedings are expected to be widely read and cited.

In addition to a regular technical program, there will be tutorials covering the state-of-the-art of this domain, a doctoral consortium, and an industrial program comprising of keynote speakers and practice/industry-paper tracks.