proceedings

MDPI

Extended Abstract

When Diversity Met Accuracy: A Story of Recommender Systems †

Alfonso Landin *, Eva Suárez-García and Daniel Valcarce

Department of Computer Science, University of A Coruña, 15071 A Coruña, Spain; eva.suarez.garcia@udc.es (E.S.-G.); daniel.valcarce@udc.es (D.V.)

* Correspondence: alfonso.landin@udc.es; Tel.: +34-881-01-1276

† Presented at the XoveTIC Congress, A Coruña, Spain, 27–28 September 2018.

Published: 14 September 2018

Abstract: Diversity and accuracy are frequently considered as two irreconcilable goals in the field of Recommender Systems. In this paper, we study different approaches to recommendation, based on collaborative filtering, which intend to improve both sides of this trade-off. We performed a battery of experiments measuring precision, diversity and novelty on different algorithms. We show that some of these approaches are able to improve the results in all the metrics with respect to classical collaborative filtering algorithms, proving to be both more accurate and more diverse. Moreover, we show how some of these techniques can be tuned easily to favour one side of this trade-off over the other, based on user desires or business objectives, by simply adjusting some of their parameters.

Keywords: recommender systems; collaborative filtering; diversity; novelty

1. Introduction

Over the years the user experience with different services has shifted from a proactive approach, where the user actively look for content, to one where the user is more passive and content is suggested to her by the service. This has been possible due to the advance in the field of recommender systems (RS), making it possible to make better suggestions to the users, personalized to their preferences.

Most of the research on the field focuses on the accuracy as the main objective of the systems. For example, the Netflix Prize goal was to improve the accuracy of Cinematch (Netflix recommendation system) by 10%, measured by the root mean squared error of the predictions. This competition fuelled the research and several advances came from it. However, in the wake of the results, studies have proven the inadequacy of this measure when it comes to the top-n recommendation task [1], introducing the use of IR metrics, such as precision or the normalized discounted cumulative gain (nDCG), to assess the performance of the system. To introduce these measures non-rated items are considered as non relevant. It has been acknowledged that making this consideration may underestimate the true metric value; however, it provides a better estimation of the recommender quality [2].

Other studies have also pointed out the convenience of measuring different properties of recommender systems such as diversity or novelty [3,4]. A system that is able to produce novel recommendations increases the probability of suggesting items to a user that would not have discovered by herself; this property is called serendipity. This quality is often associated with user satisfaction [5], but it is difficult to measure, usually involving online experiments. We use novelty as a proxy to measure this property. Being able to produce diverse recommendations, that make use of the full catalogue of items instead of focusing on the more popular ones, is usually an added benefit to a recommender system. Diversity is highly appreciated by vendors [6,7].

We analysed the performance of a couple of memory-based recommender systems, both using four different clustering techniques to compute the neighbourhoods. This performance was evaluated

in term of precision, diversity and novelty metrics. We also analysed how the systems perform with different values of their parameters, with the intent of showing how the performance of the systems with respect to the trade-off between accuracy and diversity/novelty can be tuned to suit the needs of the user or the business objectives.

2. Materials and Methods

We conducted a series of experiments in order to analyse the trade-off between accuracy, diversity and novelty in Recommender Systems.

2.1. Algorithms

We choose two memory-based based algorithms to analyse their performance. The first one, Weighted Sum Recommender (WSR), is a formulation of the classic user based recommender that stands out for its simplicity and performance [8]. The second one is an adaptation of Relevance-based Language Model (frequently abbreviated as Relevance Models or RM), used in text retrieval to perform pseudo relevance feedback [9]. In particular, we used the RM2 approach, which showed superior performance than RM1 [10].

Both algorithms use the notion of the neighbourhood of a user to perform their calculations. Intuitively, they decide to recommend or not an item based on the preferences of other users that are considered similar to the active one. We explored four clustering techniques to calculate these neighbourhoods with both algorithms. The first one, *k*-Nearest Neighbours (*k*-NN), is a well-known technique commonly used with neighbourhood based algorithms [11]. As a second method, we also tested a modification of the *k*-NN technique, inverted nearest neighbours (*k*-iNN), that claim to improve both novelty and accuracy [12]. Another technique we used was Posterior Probabilistic Clustering [13], in particular the model that uses the K-L divergence cost function (PPC2). Lastly, we used the Normalized Cuts (NC), a technique used in image segmentation [14], adapted to partition users into clusters. These last two techniques are hard clustering techniques, where a user can only be part of a single cluster. On the contrary, the first two are soft clustering techniques, meaning that a user can be in more than one cluster at the same time. These two methods also make use of a similarity measure, that has to be defined independently. For our research, we used the cosine similarity in both cases.

2.2. Evaluation Protocol

We report out result only on the MovieLens 100k dataset, given the space constraints, although similar trends have been observed in other collections. This is a very popular public dataset for evaluating collaborative filtering methods. It contains 100,000 ratings that 943 users gave to 1682 items. We used the splits provided by the collection to perform 5-fold cross-evaluation.

To evaluate de effectiveness of the recommendations we used the Normalized Discounted Cumulative Gain (nDCG), using the standard formulation as described in [15] with ratings as graded relevance judgements. In our experiments, only items with a rating of 4.0 or higher are considered relevant when evaluating. To assess the diversity of the recommendations we use the inverse of the Gini index [6]. When a value of the index is 0 it signifies that a single item is being recommended to all users. A value of 1 means that all items are recommended equally to all the users. To evaluate the novelty we use the mean self-information (MSI) [16]. All the metrics are evaluated at a cut-off of 10. We do this because we are interested in evaluating the quality of the top recommendations.

3. Results

We tested all the combinations of recommender and clustering techniques. For the soft clustering methods (*k*-NN and *k*-iNN) we varied the number of neighbours between 25 and 200. For the hard clustering techniques (PPC2 and NC) we obtained the results modifying the number of clusters

between 10 and 100. The results in terms of accuracy (nDCG), diversity (Gini) and novelty (MSI) can be observed in Figure 1.

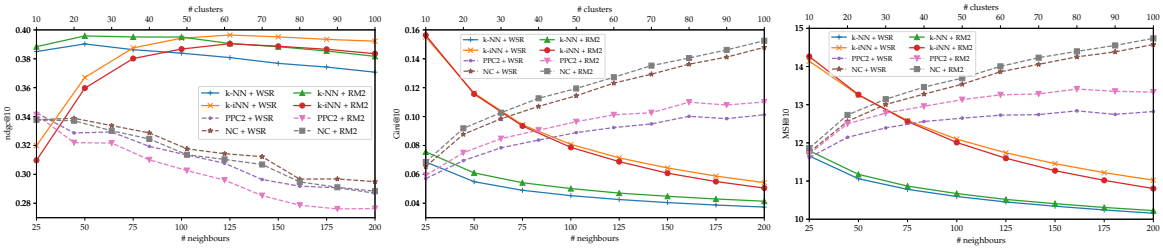

Figure 1. Values of nDCG@10, Gini@10 and MSI@10 of all studied algorithms when varying the number of clusters or neighbours.

When it comes to accuracy alone both *k*-NN and *k*-iNN show a superior performance when compared to the hard clustering methods, offering both similar results in term of nDCG. For these the type of recommender that offers the best results varies. *k*-NN obtains better results with the RM2 algorithm. In the case of *k*-iNN, it is the WSR algorithm that gets the better results.

In the case of the diversity and novelty results, it can be observed that most of the time tuning a method to provide more accurate results leads to a decrease in these other to measures. This is not always true, as can be seen with the soft clustering techniques, when increasing the numbers of neighbours too much leads to decreases in accuracy, diversity and novelty. It can also be seen that different algorithms can obtain different levels of diversity and novelty at the same level of accuracy. In this regard, the *k*-iNN method shows superior levels of diversity and novelty when compared to the *k*-NN technique at similar levels of accuracy, confirming the claim of their proponents.

4. Discussion

Results show that the intuition that during the process of tuning a recommender raising the accuracy leads to decreases in novelty and diversity holds most of the time, but there can be situations when this is no longer true, and the performance of the system moves in the same direction for all the metrics when changing a parameter.

But the results also show that the choice of algorithms is important when it comes to improving the properties of the system. It is possible to improve the performance of the system in diversity and novelty, while maintaining similar levels of accuracy. It is also possible to tune the system to balance how well it performs in all the metrics. This is a multi-objective problem and a trade off must be chosen, either by a priori setting the weight that each measure has, or by choosing any of the possible combination of parameters from the values in the Pareto front.

Funding: This work has received financial support from project TIN2015-64282-R (MINECO/ERDF), project GPC ED431B 2016/035 (Xunta de Galicia) and accreditation ED431G/01 (Xunta de Galicia/ERDF). The first author also acknowledges the support of grant FPU17/03210 (MECD). The third author also acknowledges the support of grant FPU014/01724 (MECD).

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Cremonesi, P.; Koren, Y.; Turrin, R. Performance of Recommender Algorithms on Top-N Recommendation Tasks. In Proceedings of the 4th ACM Conference on Recommender Systems (RecSys'10), Barcelona, Spain, 26–30 September 2010; ACM: New York, NY, USA, 2010; pp. 39–46, doi:10.1145/1864708.1864721.

2. McLaughlin, M.R.; Herlocker, J.L. A collaborative filtering algorithm and evaluation metric that accurately model the user experience. In Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval (SIGIR'04), Sheffield, UK, 25–29 July 2004; ACM Press: New York, NY, USA, 2004; p. 329, doi:10.1145/1008992.1009050.
3. Herlocker, J.L.; Konstan, J.A.; Terveen, L.G.; Riedl, J.T. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.* **2004**, *22*, 5–53, doi:10.1145/963770.963772.
4. McNee, S.M.; Riedl, J.; Konstan, J.A. Being Accurate is Not Enough: How Accuracy Metrics have hurt Recommender Systems. In Proceedings of the CHI'06 Extended Abstracts on Human Factors in Computing Systems (CHI EA'06), Montréal, QC, Canada, 22–27 April 2006; ACM Press: New York, NY, USA, 2006; p. 1097, doi:10.1145/1125451.1125659.
5. Ge, M.; Delgado-Battenfeld, C.; Jannach, D. Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity. In Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys'10), Barcelona, Spain, 26–30 September 2010; pp. 257–260, doi:10.1145/1864708.1864761.
6. Fleder, D.; Hosanagar, K. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Manag. Sci.* **2009**, *55*, 697–712, doi:10.1287/mnsc.1080.0974.
7. Valcarce, D.; Parapar, J.; Álvaro Barreiro. Item-based relevance modelling of recommendations for getting rid of long tail products. *Knowl.-Based Syst.* **2016**, *103*, 41–51, doi:10.1016/j.knosys.2016.03.021.
8. Valcarce, D.; Parapar, J.; Barreiro, A. Efficient Pseudo-Relevance Feedback Methods for Collaborative Filtering Recommendation. In Proceedings of the European Conference on Information Retrieval (ECIR'16), Padua, Italy, 20–23 March 2016; pp. 602–613, doi:10.1007/978-3-319-30671-1_44.
9. Lavrenko, V.; Croft, W.B. Relevance based language models. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01), New Orleans, LA, USA, 9–12 September 2001; ACM Press: New York, NY, USA, 2001; pp. 120–127, doi:10.1145/383952.383972.
10. Parapar, J.; Bellogín, A.; Castells, P.; Barreiro, A. Relevance-based language modelling for recommender systems. *Inf. Process. Manag.* **2013**, *49*, 966–980, doi:10.1016/j.ipm.2013.03.001.
11. Ning, X.; Desrosiers, C.; Karypis, G. A Comprehensive Survey of Neighborhood-Based Recommendation Methods. In *Recommender Systems Handbook*, 2nd ed.; Ricci, F., Rokach, L., Shapira, B., Eds.; Springer: Boston, MA, USA, 2015; pp. 37–76, doi:10.1007/978-1-4899-7637-6_2.
12. Vargas, S.; Castells, P. Improving Sales Diversity by Recommending Users to Items. In Proceedings of the 8th ACM Conference on Recommender Systems (RecSys'14), Foster City, CA, USA, 6–10 October 2014; ACM: New York, NY, USA, 2014; pp. 145–152, doi:10.1145/2645710.2645744.
13. Zhang, Z.Y.; Li, T.; Ding, C.; Tang, J. An NMF-framework for Unifying Posterior Probabilistic Clustering and Probabilistic Latent Semantic Indexing. *Commun. Stat.-Theory Methods* **2014**, *43*, 4011–4024, doi:0.1080/03610926.2012.714034.
14. Shi, J.; Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal.* **2000**, *22*, 888–905, doi:10.1109/34.868688.
15. Wang, Y.; Wang, L.; Li, Y.; He, D.; Chen, W.; Liu, T.Y. A Theoretical Analysis of NDCG Ranking Measures. In Proceedings of the 26th Annual Conference on Learning Theory (COLT'13), Princeton, NJ, USA, 12–14 June 2013; pp. 1–30.
16. Zhou, T.; Kuscsik, Z.; Liu, J.G.; Medo, M.; Wakeling, J.R.; Zhang, Y.C. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 4511–4515, doi:10.1073/pnas.1000488107.