Aspect-based active learning for user preference elicitation in recommender systems

María Hernández-Rubio
maria.hernandezr@estudiante.uam.es
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, Spain

Alejandro Bellogín
alejandro.bellogin@uam.es
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, Spain

Iván Cantador
ivan.cantador@uam.es
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, Spain

ABSTRACT

Recommender systems require interactions from users to infer personal preferences about new items. Active learning techniques aim to identify those items that allow eliciting a target user's preferences more efficiently. Most of the existing techniques base their decisions on properties of the items themselves, for example according to their popularity or in terms of their influence on reducing information variance or entropy within the system. Differently to previous work, in this paper we explore a novel active learning approach focused on opinions about item aspects extracted from user reviews. We thus incorporate textual information so as to decide which items should be considered next in the user preference elicitation process. Experiments on a real-world dataset provide positive results with respect to competitive state of the art methods.

KEYWORDS

user preference elicitation, active learning, user reviews, opinion mining, recommender systems

1 INTRODUCTION

Recommender Systems (RS) and, in particular, Collaborative Filtering (CF) techniques, are widely used tools that help users to find relevant information according to their preferences [13]. Being one of the most common and successful approaches to provide personalized recommendations [8, 11], CF typically needs a user–item rating matrix, where each rating reflects the preference from certain user towards a particular item. In this context, it is paramount to know as much information as possible from the users, in particular, in the form of ratings.

To overcome the lack of user information, Active Learning (AL) strategies are used to elicit such preferences in an efficient way, so that additional ratings are acquired from the users by optimizing certain goal [3]. In general, these strategies only consider properties related to the items, such as their popularity, how diverse the received ratings are (as a proxy towards their level of controversy or uncertainty), and their influence on reducing global information variance or entropy within the system. Few and recent methods, in contrast, exploit the content features of the items, such as domain attributes and metadata [15], e.g., directors and actors of a movie, genres of a music artist, and so on. These methods have demonstrated good results in the so-called item cold-start problem, which is closely related to the final goal of active learning, since in both cases there is a need for increasing the number of known ratings.

In this sense, opinions and sentiments expressed by users about items in personal, textual reviews are valuable signals of user preferences. However, their modeling and further identification is challenging. Specific features or *aspects* (e.g., technical characteristics and components) of the reviewed items –such as the *price* for cameras and mobile phones, and the *atmosphere* for restaurants and hotels– need to be extracted for properly modeling the users' preferences. In fact, the research literature on aspect extraction is extensive and has shown the positive value of aspect opinion mining for user modeling and recommendation [7].

Despite this, to the best of our knowledge, there is no AL strategy that exploits aspect opinions, instead of numeric ratings or generic non opinion-related attributes, to guide the preference elicitation process. Addressing this research gap, in this paper we explore a novel active learning approach that elicits the next item to present a user by considering its aspect-based similarity with previously interacted items by the user. We report experiments on a real-world dataset with Amazon reviews, showing that the proposed aspect-based AL strategy is able to elicit a similar number of relevant preferences with respect to existing AL strategies, but much earlier in the process, which allows mitigating the cold-start problem better and faster than item-based approaches. Moreover, we show that a recommendation method exploiting the preferences elicited by our AL strategy achieves better performance, not only in terms of rating prediction metrics as measured in previous works, but also in terms of ranking metrics, which are closer to the actual user experience [10], an issue neglected in past work that evaluated AL for recommender systems [3].

2 BACKGROUND AND RELATED WORK

As mentioned before, in this work we aim to exploit item aspects for active learning. For this purpose, we first need to extract the aspects from textual reviews provided by users. As presented in [7], aspect extraction methods can be categorized as approaches that are based on aspect vocabularies, word frequencies, syntactic relations, and topic models. Representing a starting point of our research on aspect-based AL, we focus on approaches based on aspect vocabularies, where explicit mappings between terms and aspects are specified. The use of other types of aspect extraction approaches is left as future work.

Regarding the active learning strategies used in recommender systems, the literature is extensive, so we will follow the survey and taxonomy presented in [3]. In that work, AL strategies are classified as non-personalized and personalized depending on whether they request all users for the same set of items or not. Moreover, within such categories, the authors further classified the strategies

into single- and combined-heuristic methods, which correspond to whether they implement a unique item selection rule or if several rules are somehow combined. These heuristics aim to optimize different criteria, such as reducing the uncertainty or the error in rating prediction, focusing on the items that received the highest attention from users, those more familiar to the user (and, hence, more rateable), or those that would provide the most impact on the system as a whole.

In the experiments reported in this paper, we used a representative set of strategies that cover distinct heuristics and hypotheses regarding the optimal elicited items, mostly from the non-personalized category, since they have shown a good trade-off between performance and efficiency.

3 A NOVEL ASPECT-BASED ACTIVE LEARNING METHOD

As already discussed, active learning strategies have shown promising results to elicit information about user preferences in different domains. However, when dealing with situations where users express their opinion on items by writing reviews, it becomes more important to understand and process such textual content in detail to better model the user preferences. An approach that has recently brought positive results is exploiting the rich information elements that can be extracted from the reviews, in particular, the item aspects mentioned and the opinion or sentiment associated to them.

Following a recent work on this topic [7], we propose to extract these aspects and use them to elicit those items that are more similar to the ones previously assessed by the user. Our goal is helping the user to find items that share characteristics with previously interacted items, and supporting the system to gather more preferences and better user models. Exploiting item aspects, instead of other content or collaborative information, should alleviate the cold-start problem [12], and could help the user in expressing her preferences more easily, as well as reducing the mistakes made by the recommender systems. To test this hypothesis, we leave as future work the integration and evaluation of our method within a conversational agent [1].

The proposed method for AL selects those items with higher similarities to the user's previously rated items. More specifically, we extend the item-to-item similarity matrix with the rating information already available in the system by means of a hybrid recommendation approach recently presented in [5], where a latent space is learnt based on collaborative information and side similarity, in our case, an aspect-based item similarity. We use cosine similarity over the item profile, that is, $i_n = \{w_{na}\}_{a=1}^{K}$ built on the K aspect opinions extracted for each item, where w_{na} is the weight assigned to aspect a for item i_n:

$$\text{sim}(i_n, i_m) = \frac{\sum w_{na} w_{ma}}{\sqrt{\sum w_{na}^2 \sum w_{ma}^2}} \tag{1}$$

According to the taxonomy presented in [3], this method belongs to the item-item category, and it is a personalized single-heuristic strategy, where the item-to-item similarity is computed based on the aspects extracted for each item in the system.

Regarding the exploited aspects, in this work we use a vocabulary-based aspect extraction method, in particular, the one called voc in [7]. This method makes use of a vocabulary for item aspects on a given domain, and analyzes syntactic relations between the words of each sentence in user reviews to extract the personal opinions about the aspects. We have chosen this method because it exhibits a good trade-off between simplicity and positive results. Other aspect extraction methods could be explored in the future.

4 EXPERIMENTAL SETTINGS

4.1 Dataset

As done by other researchers, in our experiments, we used the popular McAuley's Amazon product reviews dataset [9], and more specifically, we preliminary focused the experiments on the subset associated to the Movies & TV domain. Initially, this dataset includes 1,697,533 ratings, by 123,960 users on 50,052 items. Once we filtered out those items without aspects, we obtained 1,683,190 ratings on 48,074 items. Then, we decided to filter out users with less than 20 ratings since, as we shall explain in Section 4.2, the evaluation methodology is quite exhaustive, and we need to have as many users with enough information as possible. Nonetheless, in the future, we would like to extend this analysis to more formal methodologies focused on the new user problem, such as the one used in [4]. After the above process, 819,148 ratings by 14,010 users on 47,506 items remained.

Regarding the aspect coverage, the method voc introduced before initially provides 369,175 aspect annotations of 23 distinct aspects on 48,074 items, which were reduced to 367,750 aspect annotations of 23 distinct aspects on 47,506 items after the whole filtering process.

4.2 Evaluation methodology

To study the performance of the AL strategies considered in our experiments, we use the following simulation procedure, adapted from [2]. Differently to previous works where improvements were typically measured only for the same user, our procedure is oriented to evaluate the overall performance of the system.

Specifically, we divided the full dataset into 3 splits, namely *training set*, with ratings known by the system, *candidate set*, with ratings known by the users but not by the system, and *test set*, with a portion of the ratings known by the users that are withheld from the previous set. Then, a recommender system was trained using the entire training set, and for each user, in each evaluation iteration, 10 items were elicited from a particular AL strategy (which should return 10 or less items). All the items that belong to the candidate set for each user were included in the training set to be used in the next iteration. This process is repeated 170 times (iterations). At the end of each iteration, the evaluation metrics presented in the next section were computed for the recommendations generated from the updated training set.

In the experiments, we started with 2% of the data for training, 68% for candidates, and the remainder for test. Due to the high computational cost of the presented methodology, we sampled 1,500 users, and only used the above ratings from them, resulting

http://jmcauley.ucsd.edu/data/amazon/

in 80K-90K ratings on around 27K items, on average. Additionally, we repeated this procedure 3 times (where the splitting was done randomly) to report average metric values.

4.3 Evaluation metrics

In the last years, performance evaluation of recommender systems has shifted from measuring rating prediction errors to measuring ranking quality [6]. However, since most of the literature on preference elicitation used error metrics, such as Mean Absolute Error (MAE), we decided to report this metric, and thus better compare our proposal against the most influential research works.

We also report ranking-based metrics such as Precision at different cutoffs, by following the `RelPlusN` methodology described in [14], where a ranking is created for every user's relevant item where that item, together with other N items ($N = 100$ in our experiments) randomly selected, are ranked according to the estimated scores by the recommender. In this procedure, we assume items are relevant for a user whenever the rating in the test set is 5.

4.4 Baseline methods

Regarding the AL strategies compared with our method presented in Section 3, we consider the following baselines: `random`, `variance` –which selects the items with the highest variance–, `popularity` –which selects the most popular items–, `entropy` –which selects the items with the highest dispersion of the ratings for an item–, and `log-pop-entropy` –which finds a balance between the last two previous strategies. We also considered a wide array of non-personalized methods because in the original paper of the item-item approach [12], its performance was not always superior to other non-personalized strategies.

5 RESULTS

Figure 1 shows the evolution on the number of ratings correctly elicited by each strategy, that is, how many elicited items belong to the candidates set in each iteration. We observe that most of the strategies converge to the same number except our proposal, which has a limited item coverage since not all items have aspect opinions. However, we want to emphasize an interesting phenomenon; in the first 25 iterations, the strategies are divided into two groups: aspects and random (which provide less correct items), and the remainder. This observation is important when considered in combination with the performance of the recommender.

Indeed, Figure 2 shows the evolution on the error of the different AL strategies. Again, we observe that, on the long run, ours is the worst performing method (not shown for space constraints). However, in the first iterations, it is the strategy that reduces the most the error of the system. This by itself is remarkable, since it is the most useful scenario for a real user, who does not want to spend 50 iterations giving feedback to the system. This result is even more positive considering that the number of elicited preferences is smaller than with other strategies. Hence, the proposed method is able to reduce the error of the system during the first iterations of the preference elicitation process, even though it is not as competitive as other methods when finding the known items that were hidden to the system, i.e., the items in the candidate set.

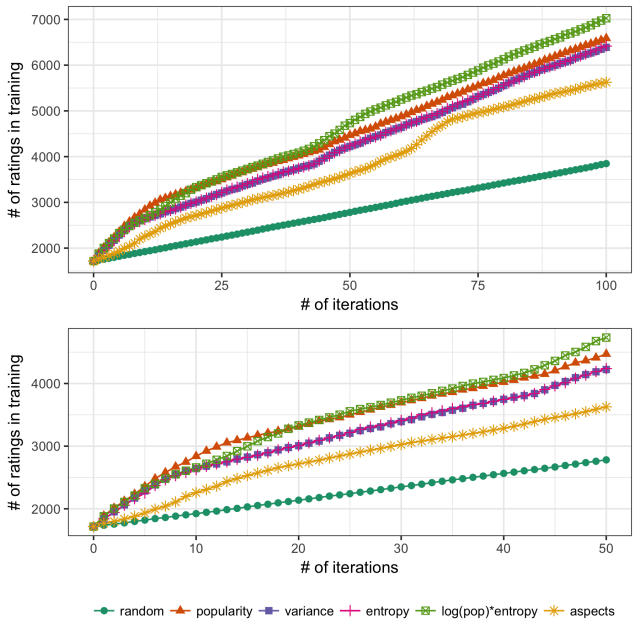

Figure 1: Evolution on the number of ratings correctly elicited by each strategy (zoomed in on the first 50 iterations in the bottom figure).

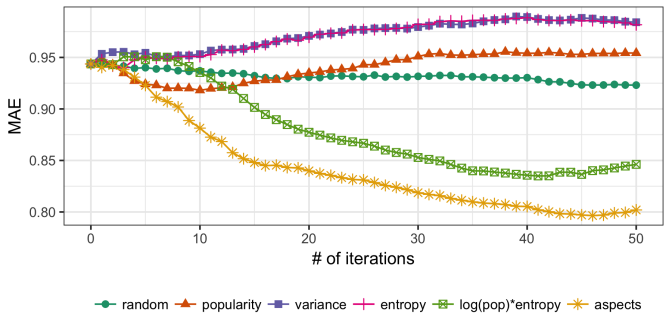

Figure 2: Evolution on the error accuracy (the lower, the better) under the effect of six elicitation strategies.

As a consequence, these results evidence that the elicited items are of higher quality with respect to those retrieved by the baselines.

At the top of Figure 3, we show the temporal evolution of the precision values on each iteration. To better expose the behavior of the metric, we also include (at the bottom of the figure) a smoothed version where the presented value is the average of the last 3 points. We observe, more clearly in this second figure, how the proposed strategy is the best performing method throughout most of the elicitation processes, also in terms of ranking metrics, in agreement with the results obtained for error metrics.

6 CONCLUSIONS

In this paper, we have proposed a novel active learning approach based on opinions about item aspects. We have preliminary shown its effect on user preference elicitation by experimenting with a

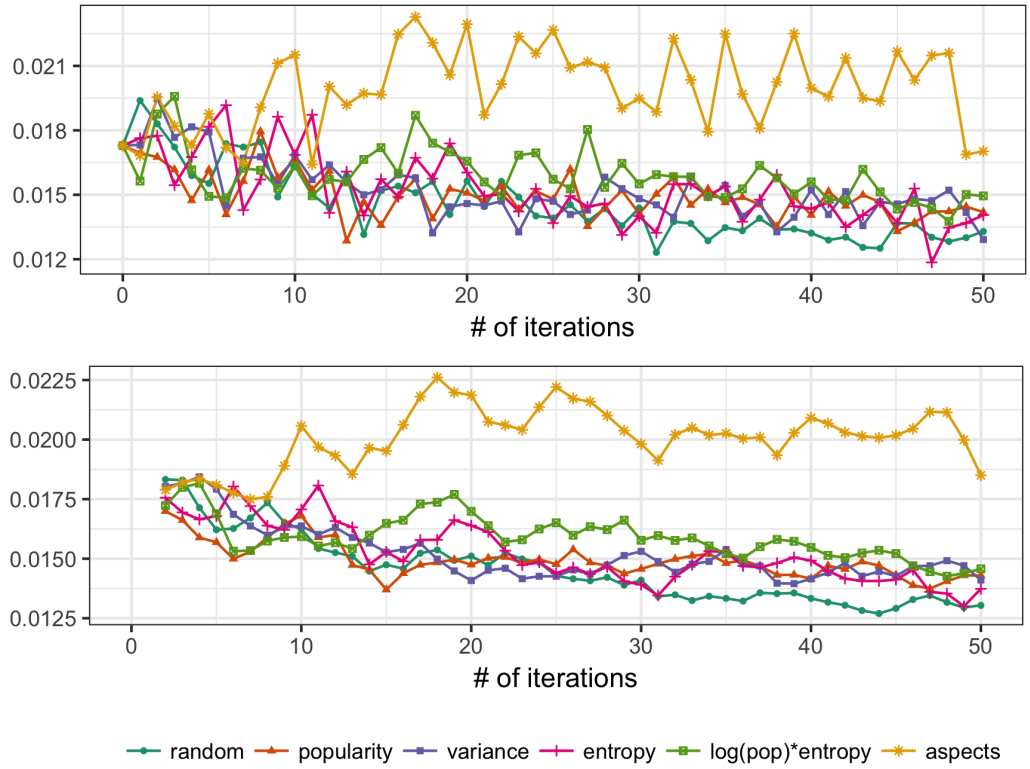

Figure 3: Ranking accuracy measured as P@5 (the higher, the better) under the effect of six elicitation strategies. Top figure: evolution of the metric value on different iterations; bottom figure: smoothed values taking the average of the last 3 points.

real-world dataset. In our empirical results, the developed method outperformed state-of-the-art strategies in terms of both rating prediction error and ranking precision metrics, computed after a recommender system was trained with user preferences elicited by each of the active learning strategies.

These results are very promising, even though we took many simple solutions to address some of the issues at hand. In this sense, in the future we aim to consider more exhaustive experiments testing several recommender systems, more sophisticated aspect extraction methods than the one used here, and datasets from several domains with different characteristics [7]. Additionally, we plan to formally analyze the behavior of our method on different cold-start settings [4], together with an online evaluation with real users, for instance, by integrating our method into a conversational agent or chatbot, with which we will check whether the user preferences are elicited faster or with higher quality, as we have observed in the offline experiments herein presented.

ACKNOWLEDGMENTS

This work was supported by the Spanish Ministry of Science and Innovation (PID2019-108965GB-I00).

REFERENCES

[1] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems. In *KDD*. ACM, 815–824.

[2] Mehdi Elahi, Francesco Ricci, and Neil Rubens. 2013. Active learning strategies for rating elicitation in collaborative filtering: A system-wide perspective. *ACM Trans. Intell. Syst. Technol.* 5, 1 (2013), 13:1–13:33.

[3] Mehdi Elahi, Francesco Ricci, and Neil Rubens. 2016. A survey of active learning in collaborative filtering recommender systems. *Comput. Sci. Rev.* 20 (2016), 29–50.

[4] Ignacio Fernández-Tobías, Iván Cantador, Paolo Tomeo, Vito Walter Anelli, and Tommaso Di Noia. 2019. Addressing the user cold start with cross-domain collaborative filtering: exploiting item metadata in matrix factorization. *User Model. User-Adapt. Interact.* 29, 2 (2019), 443–486.

[5] Evgeny Frolov and Ivan V. Oseledets. 2019. HybridSVD: when collaborative information is not enough. In *RecSys*. ACM, 331–339.

[6] Asela Gunawardana and Guy Shani. 2015. Evaluating Recommender Systems. In *Recommender Systems Handbook*. Springer, 265–308.

[7] María Hernández-Rubio, Iván Cantador, and Alejandro Bellogín. 2019. A comparative analysis of recommender systems based on item aspect opinions extracted from user reviews. *User Model. User-Adapt. Interact.* 29, 2 (2019), 381–441.

[8] Yehuda Koren and Robert M. Bell. 2015. Advances in Collaborative Filtering. In *Recommender Systems Handbook*. Springer, 77–118.

[9] Julian J. McAuley and Alex Yang. 2016. Addressing Complex and Subjective Product-Related Queries with Customer Reviews. In *WWW*. ACM, 625–635.

[10] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI Extended Abstracts*. ACM, 1097–1101.

[11] Xia Ning, Christian Desrosiers, and George Karypis. 2015. A Comprehensive Survey of Neighborhood-Based Recommendation Methods. In *Recommender Systems Handbook*. Springer, 37–76.

[12] Al Mamunur Rashid, István Albert, Dan Cosley, Shyong K. Lam, Sean M. McNee, Joseph A. Konstan, and John Riedl. 2002. Getting to know you: learning new user preferences in recommender systems. In *IUI*. ACM, 127–134.

[13] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. Recommender Systems: Introduction and Challenges. In *Recommender Systems Handbook*. Springer, 1–34.

[14] Alan Said and Alejandro Bellogín. 2014. Comparative recommender system evaluation: benchmarking recommendation frameworks. In *RecSys*. ACM, 129–136.

[15] Yu Zhu, Jinghao Lin, Shibi He, Beidou Wang, Ziyu Guan, Haifeng Liu, and Deng Cai. 2020. Addressing the Item Cold-Start Problem by Attribute-Driven Active Learning. *IEEE Trans. Knowl. Data Eng.* 32, 4 (2020), 631–644.