Recommender Systems from an Industrial and Ethical Perspective

Dimitris Paraschakis
Dept. of Computer Science
Malmö University
SE-205 06 Malmö, Sweden
dimitris.paraschakis@mah.se

ABSTRACT

Over the recent years, a plethora of recommender systems (RS) have been proposed by academics. The degree of adoptability of these algorithms by *industrial* e-commerce platforms remains unclear. To get an insight into real-world recommendation engines, we survey more than 30 existing shopping cart solutions and compare the performance of popular recommendation algorithms on proprietary e-commerce datasets. Our results show that deployed systems rarely go beyond trivial "best seller" lists or very basic personalized recommendation algorithms, which nevertheless exhibit superior performance to more elaborate techniques both in our experiments and other related studies. We also perform chronological dataset splits to demonstrate the importance of preserving the sequence of events during evaluation, and the recency of events during training. The second part of our research is still ongoing and focuses on various *ethical* challenges that complicate the design of recommender systems. We believe that this direction of research remains mostly neglected despite its increasing impact on RS' quality and safety.

Keywords

industrial recommender systems; recommendation ethics; Netflix Prize; ethical recommendation framework; e-commerce; recommender systems survey

1. INTRODUCTION: ECHOES FROM THE NETFLIX PRIZE

Recommender systems emerged as an independent research area of machine learning in mid-1990s and received a great deal of attention after the announcement of the Netflix Prize[1] in 2006. The contest was set to crowdsource recommendation algorithms from the large research community, whose goal was to surpass Netflix' own algorithm by 10% in terms of accuracy (measured in RMSE). After 10 years, the Netflix Prize still teaches us lessons. One of them is the fact that the prize-winning algorithm was never put to real use: Netflix concluded that the measured accuracy gains "did not seem to justify the engineering effort needed to bring them into a production environment" [1]. This motivates the investigation of the receptiveness of *e-commerce* platforms to recent academic advances in the field of recommender systems. We trust that the high *practical* value of recommender systems in e-commerce ought to be the reason why this research field exists in the first place. As Pradel *et al.* [13] note, "case-studies are necessary to better understand the specificities of purchase datasets and the factors that impact recommender systems for retailers". The first part of our work is such a case study. The specificity of purchase datasets has to do with the absence of explicit feedback (such as movie ratings) and the extreme sparsity of data (because of the severe long-tail effect). As a result, the majority of algorithms designed for *rating* datasets cannot be directly applied to *purchase* data. Furthermore, many e-commerce domains (e.g. fashion, travel, etc.) exhibit more profound seasonality effects in comparison to movie rating datasets. Therefore, we argue that the realistic offline evaluation of recommendation algorithms in industrial contexts has to be done on time-based dataset splits, as we explain in Section 2. We also present the results of the survey and our comparative study of several recommendation algorithms.

Aside from the apparent research shift towards explicit feedback RS, the comparably smaller amount of research on e-commerce RS may also be attributed to the lack of publicly available retail datasets. In our experience, many retailers are reluctant to release their sensitive purchase data because of the *failure of anonymization* [10]. This is where another motivating lesson from the Netflix Prize comes in. This time, we learn the lesson of ethics: two years after its public release, the Netflix dataset was de-anonymized via a linking attack [8], putting the privacy of 500,000 users at risk. This resulted in a lawsuit and put an end to the planned Netflix Prize sequel [7]. The problems of ethics in recommender systems certainly go far beyond anonymization and relate to areas such as data collection and filtering, algorithmic opacity and biases, behaviour manipulation, A/B testing, etc. According to the recent study by Tang & Winoto [14], there exist only two publications (apart from [14]) that specifically address the problem of ethical recommendations. Therefore, this research direction deserves further attention. We discuss ethical issues briefly in Section 3. We conclude the paper and outline our future work in Section 4.

[1] http://www.netflixprize.com/

RecSys '16, September 15 - 19, 2016, Boston , MA, USA

ISBN 978-1-4503-4035-9/16/09...$15.00

DOI: http://dx.doi.org/10.1145/2959100.2959101

2. E-COMMERCE RECOMMENDER SYSTEMS: AN INDUSTRIAL PERSPECTIVE

2.1 Methodology

Our study[2] employs mixed methodology to meet its goals. First, we conduct offline experiments to compare the performance of algorithms from three families of collaborative filtering (CF):

- MF-based CF: *Weighted Regularized Matrix Factorization (**WRMF**)*, *Bayesian Personalized Ranking Matrix Factorization (**BPRMF**)*
- Memory-based CF: *User-user K-nearest neighbours (**UserKNN**)*
- Data mining: *Association Rules Miner (**ARM**)*

Most popular and ***random*** product recommendations are used as baselines. The accuracy of algorithms is measured in terms of *MAP*, *F1@5*, and *R-precision*. The optimal parameters for algorithms are estimated via golden section search. We refer the reader to [11] for the technical details of the above algorithms and evaluation metrics. The evaluation is performed both on *random* and *chronological* dataset splits. The datasets come from two real e-commerce retailers: a fashion store (denoted as D1) and a book store (denoted as D2). The summary of the two datasets is given in Table 1. The results of the experiments are presented in Section 2.3.1.

Table 1: Summary for datasets D1 and D2

	D1	**D2**
Domain	Fashion	Books
Timespan	9 months 26 days	3 months 10 days
Customers	26,091	505,136
Products	4706	210,455
Events	78,449	1,623,576
Sparsity	99.936%	99.998%

Second, we run an online survey[3] to assess recommendation engines of existing e-commerce platforms on three criteria: a) *properties of recommendations*, b) *data utilized for recommendations*, and c) *recommendation techniques*. The selection of these techniques is based on the classification by Amatriain *et al.* [2] (with the addition of several extra techniques). The survey results are given in Section 2.3.2.

2.2 Evaluation

In our experiments, we follow two data splitting strategies:

1. **Traditional approach (random split)**. As it is often practised in the RS literature, we discard *cold-start users* (having less than 10 purchases in our case) and then split user profiles into training and test sets using random sampling (50/50 ratio in our case). Finally, we divide users into 5 folds for cross-validation.

2. **Proposed approach (chronological split)**. To attempt a more realistic evaluation, we keep *cold-start users*[4] in the dataset and divide it using time-based split points. In particular, we propose an *expanding time window* approach, where after setting the temporal split point, the recommender is trained on increasing portions of the training set, starting from the most recent events and ending with the full history.

The motive for attempting both approaches was to determine whether the commonly used random splits can reliably reproduce the ranking of algorithms coming from more realistic temporal, cold-start splits.

2.3 Results

2.3.1 Experiments

The performance of algorithms in *random* mode is given in Tables 2 and 3. It can be seen that UserKNN and ARM outperform MF-based methods on all metrics. This is particularly evident in case of D2, where BPRMF performs on the same level as the recommender of most popular items. This finding is in line with those reported in [13].

In *chronological* mode, the temporal dataset partitioning resulted in the following splits:

- D1: test set containing the last month of data and 9 training sets of increasing length from 1 to 9 months;
- D2: test set containing the last 2 weeks of data and 6 training sets of increasing length from 2 to 12 weeks;

From observing the accuracy measurements of algorithms for each training set[5], we can see that the optimal training history for D1 amounts to the last 3 months of data, which appears reasonable for a fashion store. Likewise, the trend of favouring recent events was observed in D2. It was also evident that the *best seller* list's accuracy peaked when it was trained on the most recent and smallest chunk of data. After averaging accuracy scores over all training sets (Tables 4 and 5), we noticed that the relative ranking of algorithms was very similar to that of the random split, with UserKNN and ARM being on top. In absolute terms, however, the accuracy scores dropped down significantly in the chronological mode. Whereas this can be attributed to the presence of cold-start users, in our additional experiments (not reported here) we show empirically that the random splitting itself can significantly overestimate the accuracy of algorithms.

In terms of running times, MF-based methods were either (fairly) accurate or fast, but never both. The accuracy-speed trade-off of these models is adjusted by balancing between the number of iterations and the number of factors. The best accuracy-speed ratio was achieved by ARM, which also required no parameter tuning.

2.3.2 Survey

20 commercial and 11 open-source e-commerce platforms participated in the survey, whose responses are summarized in Figure 1.

We can see that *transactional data* is used by more than 80% of respondents, which points to the clear dominance of CF over content-based filtering. Surprisingly, *temporal data* is mostly neglected by industrial RS, which contrasts to our empirical findings. It seems that commercial platforms tend to put more effort in *personalization* than their open-source rivals. We observe that *neighborhood-based CF* and *association rules mining* are the most popular personalized recommenders, which is totally in line with our experimental results. Recommending *best sellers* is the preferred approach for most platforms. Indeed, our experiments show that outperforming best-seller lists can be a challenging task even for personalized recommenders. Moreover, trivial *random* and *manual* recommenders appear more widespread than state-of-the-art techniques, such as MF-based algorithms.

[2]published in full in [11]

[3]http://goo.gl/forms/zFEfLLlFHO

[4]those with at least 2 events to allow per-user splitting

[5]the performance charts are omitted due to limited space

Table 2: Scores for the random split on D1

Recommender	*MAP*	*F1@5*	*R-prec*	*Running Time*
WRMF	0.0927	0.1013	0.1036	00:05:21
BPRMF	0.0650	0.0676	0.0698	00:00:32
UserKNN	0.1000	0.1078	0.1073	00:00:05
ARM	**0.1100**	**0.1162**	**0.1155**	00:00:01
Most Popular	0.0523	0.0458	0.0468	00:00:00
Random	0.0028	0.0000	0.0010	00:00:00

Table 3: Scores for the random split on D2

Recommender	*MAP*	*F1@5*	*R-prec*	*Running Time*
WRMF	0.1293	0.1322	0.1363	07:10:10
BPRMF	0.0188	0.0148	0.0187	00:12:46
UserKNN	**0.1796**	**0.1859**	**0.1854**	00:56:21
ARM	0.1774	0.1813	0.1825	00:16:24
Most Popular	0.0139	0.0149	0.0156	00:07:02
Random	0.0002	0.0001	0.0001	00:09:09

Table 4: Average scores for the chronological split on D1

Recommender	*MAP*	*F1@5*	*R-prec*	*Running Time*
WRMF	0.0244	0.0188	0.0147	00:05:27
BPRMF	0.0214	0.0121	0.0095	00:00:06
UserKNN	0.0312	**0.0229**	0.0222	00:00:06
ARM	**0.0334**	0.0214	**0.0225**	00:00:01
Most Popular	0.0206	0.0113	0.0080	00:00:00
Random	0.0037	0.0014	0.0009	00:00:00

Table 5: Average scores for the chronological split on D2

Recommender	*MAP*	*F1@5*	*R-prec*	*Running Time*
WRMF	0.0351	0.0260	0.0224	02:27:10
BPRMF	0.0160	0.0130	0.0114	00:18:37
UserKNN	0.0469	0.0363	**0.0352**	00:50:14
ARM	**0.0488**	**0.0377**	0.0346	00:20:35
Most Popular	0.0150	0.0114	0.0094	00:10:24
Random	0.0001	0.0000	0.0000	00:14:42

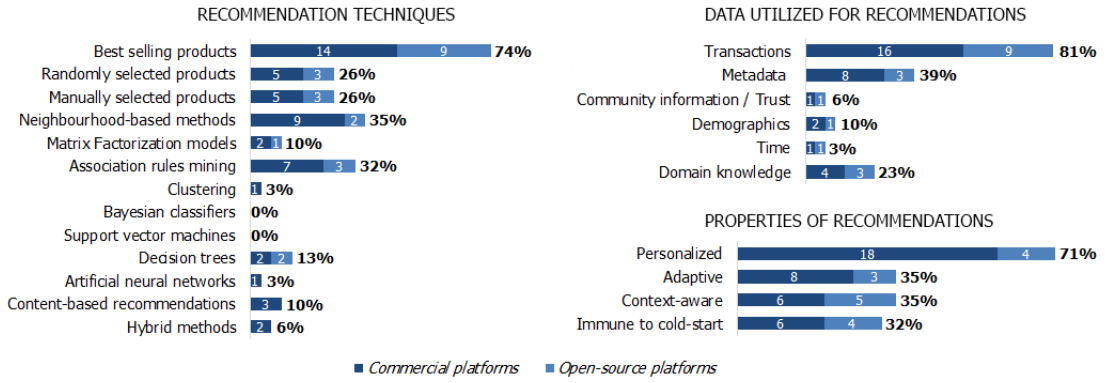

Figure 1: Survey responses from 31 e-commerce platforms

3. TOWARDS ETHICAL RECOMMENDER SYSTEMS

Our ongoing work focuses on another rather overlooked aspect of recommender systems, which is the problem of *ethics*. This problem is multifaceted and relates to many broader areas of data science. In this section, we give a brief overview of some ethical challenges in and around recommender systems.

Data collection. The lack of *transparency* and *informed consent* is what explains the great demand in "do not track" tools that would help users gain control over the data collection process. Furthermore, enriching user profiles by means of tracking cookies, linked open data, social networks, and even data brokers increase the risk of user privacy breaches.

User profiling. User behaviour profiles built for serving personalized recommendations may as well be utilised for malicious purposes, such as *phishing* or *social engineering* [7]. Moreover, disclosed user profiles may reveal sensitive private information. *Profile injection* is another possible attack that uses fake profiles to promote or demote recommendations of certain items [3], e.g. by artificially influencing their ratings. A number of *privacy preserving collaborative filtering* (PPCF) algorithms have been proposed in RS literature (e.g. [15]) to protect user profiles from leakage. The major challenge in such systems is to preserve recommendation accuracy.

Data publishing. The massive research on RS would not be possible without publicly released datasets (MovieLens, Netflix, etc.), which have positive impact on both research and practice [5]. Because a *public* dataset typically contains *private* data, releasing a myriad of user records is a serious and responsible moral act. The naive assumption that the dataset remains safe as long as *personally identifiable information* has been disguised has long ceased to hold. Many examples of user re-identification in anonymized datasets (e.g. [9]) have been reported in the literature (recall the Netflix case). The existence of quasi-identifiers and rich "outside" information makes de-anonymization of public datasets a persistent threat. On the other hand, an attempt of aggressive anonymization can render the dataset useless for a RS.

Data filtering. With very few exceptions (e.g. [14]), contemporary RS do not employ any moral filtering of their output. This raises an issue of content censorship with all its consequent implications. As noted in [14], what is *algorithmically* appropriate is not necessarily *ethically* appropriate. Should the RS be held liable for serving offensive or hazardous content? How to balance between commercial and moral values in a RS? The ethical appropriateness of candidate items can be established by mapping potentially harmful elements in media content (drug use, nudity, etc.) to a user's persona (gender, age, etc.), as explained in [14]. Authors suggest that moral values are incorporated in system requirements, with users having control over the filtering process.

Algorithmic opacity. A typical RS operates as a "black box", with its *output* being the only part that is visible to a user. Collecting data and processing it into recommendations is done completely behind the scenes. What are the ethical implications of not knowing the *algorithm* that rules recommendations? Apparently, this opacity makes it hard to reason about the political, economic, and cultural agendas behind these suggestions [12]. From the corporate ethics perspective, however, the intentional secrecy is needed to retain company's competitive advantage. Even when an algorithm is not kept in secrecy, explaining it to a user might be hardly possible because of its inherent mathematical complexity.

Biases and behaviour manipulation. In their interaction with RS, how can users be certain that their interests are respected and prioritized? What if the users' behaviour is being algorithmically manipulated to meet RS's own objectives (e.g. selling expensive / overstocked / nearly expired products)? One example of an unethical algorithmic bias is *price discrimination*, i.e. charging customers different prices based on their perceived willingness to pay [4]. How to ensure that the RS is unbiased? One way to aid transparency is to implement an *explanation interface* [6] for recommendations.

A/B testing. Many industrial RS resort to *A/B testing* to assess the effects of tweaking their algorithms. In most cases, visitors are silently dragged into these experiments without their knowledge, let alone consent. Should an e-commerce interface allow users to opt out of A/B testing? Is it their moral right to demand interacting with the "real" recommender system instead of the one being put to test?

4. CONCLUSIONS AND FUTURE WORK

In this paper, we have taken a closer look at *industrial* recommender systems in e-commerce and have touched upon certain *ethical* implications that affect the design and implementation of these systems.

After experimenting on retail data and surveying existing e-commerce platforms, we conclude that their adoption of sophisticated RS that have been proposed by the research community over the years is rather slow. Whether it has to do with the unjustifiable "engineering efforts" that had stopped Netflix from implementing a better algorithm, or strict real-time performance requirements of industrial shopping cart solutions, the suitability of these models to the realities of the e-commerce realm requires further investigation.

Our experiments also illustrate the importance of *time-aware* evaluation, where dataset splits are done chronologically and the sequence of events in user profiles is preserved. Without this property, the analysis of a retail dataset is prone to significant overestimation of the algorithm performance due to capturing erroneous shopping patterns. We also show that training on all available history is in most cases suboptimal. The proposed *expanding time window* method can be used to establish the optimal timespan of the training set, thus improving recommendation accuracy and speed.

Finally, we observe that multiple moral dilemmas of varied severity emerge on virtually every stage of RS design. We plan to continue our work on identifying their impact and possible solutions in hope to outline an ethical framework for RS designers.

5. REFERENCES

[1] X. Amatriain and J. Basilico. Netflix recommendations: Beyond the 5 stars. Online: http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html/, Apr. 2012. Accessed: 2015-03-12.

[2] X. Amatriain, A. Jaimes, N. Oliver, and J. Pujol. Data mining methods for recommender systems. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, chapter 2, pages 257–297. Springer-Verlag New York, Inc., 2010.

[3] R. Burke, M. O'Mahony, and N. Hurley. Robust collaborative recommendation. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, chapter 25, pages 805–835. Springer-Verlag New York, Inc., 2010.

[4] N. Diakopoulos. Algorithmic accountability reporting: On the investigation of black boxes. *Tow Center for Digital Journalism A Tow/Knight Brief*, 2014.

[5] F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5:19:1–19:19, Dec. 2015.

[6] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 241–250. ACM, 2000.

[7] A. Koene, E. Perez, C. J. Carter, R. Statache, S. Adolphs, C. O'Malley, T. Rodden, and D. McAuley. Ethics of personalized information filtering. In *Second International Conference on Internet Science, INSCI*, pages 123–132, 2015.

[8] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pages 111–125, 2008.

[9] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *30th IEEE Symposium on Security and Privacy*, pages 173–187. IEEE, 2009.

[10] P. Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57:1701, 2009.

[11] D. Paraschakis, B. J. Nilsson, and J. Holländer. Comparative evaluation of top-n recommenders in e-commerce: An industrial perspective. In *14th International Conference on Machine Learning and Applications (ICMLA)*, pages 1024 – 1031. IEEE, 2015.

[12] F. Pasquale. The black box society. the secret algorithms that control money and information. Harvard University Press, 2015.

[13] B. Pradel, S. Sean, J. Delporte, S. Guérif, C. Rouveirol, N. Usunier, F. Fogelman-Soulié, and F. Dufau-Joel. A case study in a recommender system based on purchase data. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 377–385, 2011.

[14] T. Tang and P. Winoto. I should not recommend it to you even if you will like it: the ethics of recommender systems. *New Review of Hypermedia and Multimedia*, 19:111–138, 2016.

[15] J. Zhan, C.-L. Hsieh, I.-C. Wang, T.-S. Hsu, C.-J. Liau, and D.-W. Wang. Privacy-preserving collaborative recommender systems. *Trans. Sys. Man Cyber Part C*, 40:472–476, July 2010.