Accuracy and Diversity in Cross-domain Recommendations for Cold-start Users with Positive-only Feedback

Ignacio Fernández-Tobías
Universidad Autónoma de Madrid
ignacio.fernandezt@uam.es

Paolo Tomeo
Polytechnic University of Bari
paolo.tomeo@poliba.it

Iván Cantador
Universidad Autónoma de Madrid
ivan.cantador@uam.es

Tommaso Di Noia
Polytechnic University of Bari
tommaso.dinoia@poliba.it

Eugenio Di Sciascio
Polytechnic University of Bari
eugenio.disciascio@poliba.it

ABSTRACT

Computing useful recommendations for cold-start users is a major challenge in the design of recommender systems, and additional data is often required to compensate the scarcity of user feedback. In this paper we address such problem in a target domain by exploiting user preferences from a related auxiliary domain. Following a rigorous methodology for cold-start, we evaluate a number of recommendation methods on a dataset with positive-only feedback in the movie and music domains, both in single and cross-domain scenarios. Comparing the methods in terms of item ranking accuracy, diversity and catalog coverage, we show that cross-domain preference data is useful to provide more accurate suggestions when user feedback in the target domain is scarce or not available at all, and may lead to more diverse recommendations depending on the target domain. Moreover, evaluating the impact of the user profile size and diversity in the source domain, we show that, in general, the quality of target recommendations increases with the size of the profile, but may deteriorate with too diverse profiles.

Keywords

Cross-domain recommendation, cold-start, diversity

1. INTRODUCTION

Providing relevant suggestions of items for new users is a well-known problem in recommender systems. In such cases there is little or no information about the users preferences, and traditional recommendation models are not able to compute meaningful personalized predictions. To compensate this lack of information, two major approaches have been studied in previous work: (i) preference elicitation techniques [10] that directly ask the user to provide some ratings before delivering recommendations, and (ii) methods that exploit additional information about the users to better estimate their preferences. In the latter case, some approaches combine content and collaborative information [12], and others exploit demographic data or even the user's personality [4] to address the user cold-start problem.

More recently, cross-domain recommender systems [1] that leverage additional information from different but related source domains have been introduced as a potential solution to cold-start situations. This auxiliary information can be exploited to mitigate the lack of historical data in the target recommendation domain, thus addressing the user cold-start [3]. In one of the first papers on the topic, Winoto and Tang [15] conjectured that although the introduction of cross-domain information could deteriorate the prediction performance in the general –non cold-start– case, it could still lead to more diverse recommendations. Subsequent work proposed methods to effectively learn and transfer knowledge from the source domain to the target [5], and found that the quality of the recommendations improves when the involved domains are semantically more related [11]. Nevertheless, to the best of our knowledge, no previous work has tested Winoto and Tang's conjecture regarding the diversity of recommendations when cross-domain data is exploited.

Moreover, in [2] it has been shown that users perceive differences in the recommendation quality depending on the variety of items they naturally prefer. Based on this observation, we hypothesize that both the amount and diversity of source domain preferences have an impact on the accuracy of cross-domain recommendations. Specifically, we identify three main research questions:

- **RQ1** ***How beneficial in terms of accuracy is to exploit cross-domain information for cold-start users?*** We analyze the ranking performance of the top-N recommendations based on positive-only feedback, following a principled evaluation methodology specifically designed for cold-start scenarios [7].
- **RQ2** ***Is cross-domain information really useful to improve the recommendation diversity?*** In order to test Winoto and Tang's conjecture we include in the evaluation the intra-list diversity metric and the recently proposed binomial diversity framework [14].
- **RQ3** ***What is the impact of the size and diversity of the user profile in the source domain on the***

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

DOI: http://dx.doi.org/10.1145/2959100.2959175

quality of the target recommendations? We check this by computing the degree of diversity of the user profiles in the source domain. This work represents, to the best of our knowledge, the first analysis on user profile diversity for cross-domain recommendation.

We investigate these issues by evaluating a number of memory-based and matrix factorization algorithms in single and cross-domain scenarios, using two datasets with positive-only only feedback consisting of Facebook *likes* on movies and music artists, mapped to DBpedia[1] entities, whose metadata is used to also evaluate two state-of-the-art graph-based methods able to exploit heterogeneous information in the recommendation process.

2. EXPERIMENTAL SETTING

Dataset. The recommendation models presented in this paper were evaluated on a Facebook dataset with user *likes* for movie and music items, which we extended with item metadata extracted from DBpedia. In [13] we detail the dataset and the developed process to automatically extract DBpedia semantic networks relating items and features. Next we provide a brief summary of them. In the original raw data –as acquired from the Facebook Graph API– each user-*like*-item relation was given as a 4-tuple with the identifier, name and category of the liked item, and the timestamp of the *like* creation, such as {id: "35481394342", name: "The Matrix", category: "Movie", created_time: "2015-05-14T12:35:08+0000"}. Distinct names may exist for the same item, e.g., "The Matrix", "The Matrix: Film series" and "The Matrix (saga)" for "The Matrix" movie saga. Users thus may provide likes for different Facebook pages referring to the same item. Consolidating and unifying the items of the extracted Facebook *likes*, our method automatically maps the items names to the unique URIs of the corresponding DBpedia entities, e.g., `http://dbpedia.org/resource/The_Matrix` for the identified names of "The Matrix" movie saga. A core stage in the method is to execute SPARQL queries to the DBpedia endpoint that (i) map item names with entity labels, expressed through the `rdfs:label` property, (ii) disambiguate entities using the `rdf:type` property and the Facebook item category field, and (iii) consider equivalent item names by means of the `dbo:wikiPageRedirects` property.

Evaluated recommendation methods. We evaluated the following recommendation algorithms in single and cross-domain scenarios, using the validation set to tune model hyperparameters in each case.

POP: Recommends the most popular items not yet liked by the user. **UNN**: User-based nearest neighbors with Jaccard similarity and neighborhood size of $k = 100$. **INN**: Item-based nearest neighbors with Jaccard similarity and indefinite neighborhood size. **IMF**: Hu et al.'s matrix factorization method for positive-only feedback [6] with 29 factors for movies and 21 factors for music.

Thanks to the linking of items to entities in the DBpedia knowledge graph, we are able to exploit algorithms that leverage the graph-based nature of the underlying side information. In particular, we built a hybrid graph as proposed in [9] and we used it as input for the following two algorithms. **HeteRec**: Graph-based recommender system proposed in [16], based on a diffusion method of user preferences following different meta-paths. **PathRank**: Personalized PageRank considering the connectivity between users and items along different meta-paths [8].

For UNN, INN, IMF, HeteRec and PathRank we considered both their application to single-domain scenarios and to cross-domain ones. Hereafter we use the prefix "CD-" to indicate the cross-domain version of the corresponding algorithm.

Evaluation methodology. For the evaluation we follow the user-based 5-fold cross-validation strategy proposed in [7] for cold-start scenarios. First, we select users in the target domain with at least 16 likes and split them into five equally sized subsets. For each fold, we keep all the data from the other folds in the training set, whereas the likes from the users in the selected fold were randomly split into three subsets: training set (10 likes), validation set (5 likes), and testing (remaining likes, hence at least 1). In order to simulate different user profile sizes from 1 to 10 likes, we repeat the training and the evaluation ten times, starting with the first like in the training set and incrementally increasing it one by one. This setting allows us to evaluate each profile size with the same test set, avoiding potential biases in the evaluation due to different test set sizes [7]. After this preprocessing, the Facebook music dataset contains $49,369$ users, $5,748$ music bands or artists, and $2,084,462$ likes; the movie dataset contains $26,943$ users, $3,901$ movies, and $876,501$ likes. The user overlap for movies is 89.96% and music is 84.69%. In order to simulate the cross-domain scenario, we simply append the full source domain dataset to the previous training set. We used the Mean Reciprocal Rank (MRR) to evaluate the ranking accuracy of the recommendations, which computes the average reciprocal rank of the first relevant item in the recommendation list. Whereas, Intra-List Diversity (ILD) and Binomial Diversity Framework (BinomDiv) [14] were used to evaluate the individual diversity, namely the degree of diversity in the recommendation lists based on item genres extracted from DBpedia. Along with accuracy, we also measured catalog coverage as the percentage of items that are recommended at least once, to better understand the differences among the compared algorithms.

3. RESULTS

In the following we discuss the outcomes of three experiments we conducted to investigate each of the research questions stated in Section 1.

Cross-domain recommendation accuracy.

To address RQ1, we compare the accuracy of the target recommendations in single-domain and cross-domain scenarios. Table 1 shows the MRR values for movies (left) and music (right) target recommendations.

Music (source)–Movies (target). CD-UNN is the most accurate method for extreme cold-start users (0 likes in target domain), CD-INN where 1 or 2 likes are provided, and CD-IMF from 3 to 10 likes. Curiously, CD-INN and CD-HeteRec using only cross-domain information are able to beat almost all the other methods even where they use target information up to 4 likes. Moreover, CD-UNN is subject to a drastic fall from 0 to 1 like, obtaining the worst accuracy among all the methods and configuration (even lower than POP). Further analysis revealed that this is due to our choice of Jaccard as user similarity metric, which we observed pro-

[1] http://dbpedia.org

Table 1: Accuracy and diversity values for different cold-start target profile sizes.

	Source – Target	Music – Movies											Movies – Music										
	Target size	0	1	2	3	4	5	6	7	8	9	10	0	1	2	3	4	5	6	7	8	9	10
MRR ($\times 10^{-3}$)	POP	290	293	295	298	299	303	304	307	310	312	315	335	337	340	343	345	347	350	352	354	357	359
	UNN		334	324	322	330	345	379	393	402	412	422		425	394	398	422	454	485	504	525	**536**	547
	CD-UNN	**383**	279	304	321	335	347	353	368	378	394	406	433	270	307	336	373	402	438	463	490	509	526
	INN		233	308	334	359	374	388	403	408	420	426		320	389	430	455	476	491	506	520	533	544
	CD-INN	347	**352**	**358**	367	369	374	382	388	392	397	403	419	**437**	**457**	**471**	**480**	**492**	**503**	**514**	**526**	**536**	545
	IMF		254	292	315	335	343	363	377	389	397	417		350	396	431	452	473	489	505	522	533	**548**
	CD-IMF	304	330	354	**370**	**378**	**387**	**400**	**410**	**424**	**428**	**439**	299	358	401	429	453	477	487	501	521	531	543
	HeteRec		320	351	360	371	376	386	389	396	402	408		361	394	424	442	467	484	499	517	526	536
	CD-HeteRec	376	345	350	356	361	364	367	370	374	381	384	**527**	431	450	461	469	477	481	488	497	503	509
	PathRank		340	345	346	352	350	354	357	361	363	367		411	416	420	426	429	433	436	442	444	449
	CD-PathRank	346	317	317	321	325	327	330	333	337	341	345	495	399	402	405	411	415	419	422	427	432	436
BinomDiv@10 ($\times 10^{-3}$)	POP	**401**	304	336	354	368	378	386	**393**	**400**	**405**	**410**	324	228	262	282	295	305	313	321	327	333	338
	UNN		360	385	**404**	**392**	**396**	**394**	**393**	393	396	395		296	332	348	347	330	317	309	305	301	297
	CD-UNN	368	**404**	**386**	376	373	372	372	374	374	377	380	296	**411**	**380**	358	347	329	322	316	311	308	305
	INN		289	308	315	321	323	327	329	332	333	337		200	213	219	223	229	231	235	236	239	240
	CD-INN	309	240	268	283	297	304	310	316	322	325	330	277	231	255	264	270	272	273	274	274	276	276
	IMF		299	320	335	344	347	355	358	363	366	368		196	217	232	241	249	253	256	260	261	265
	CD-IMF	270	231	270	289	302	315	323	328	332	338	341	248	229	254	264	271	272	276	277	277	278	278
	HeteRec		311	328	334	337	341	343	346	348	350	354		227	264	280	288	296	300	302	304	306	306
	CD-HeteRec	333	271	298	314	324	333	339	345	350	354	358	372	271	314	331	342	349	354	357	361	363	366
	PathRank		317	327	336	342	352	353	359	361	366	368		350	**380**	**395**	**404**	**410**	**413**	**416**	**419**	**421**	**422**
	CD-PathRank	336	270	294	310	320	327	334	339	345	350	355	**405**	335	367	384	394	402	408	412	415	418	419

vides unreliable scores in cold-start situations. Comparing the methods between single and cross-domain configuration, we can see that only INN and IMF can benefit from music feedback in terms of accuracy. All the other methods lose accuracy when music feedback is also considered. In terms of coverage, UNN is the only method able to benefit from music feedback: UNN reaches values from 10% to 18% and CD-UNN from 38% to 50% among the different profile sizes.

Movies–Music. CD-HeteRec yields the most accurate recommendations in the extreme cold-start scenario, while CD-INN is the best method for all the other profile sizes, even though UNN obtains close values with 8 and 9 likes, and UNN and IMF overcome CD-INN with 10 likes but with a not relevant difference. CD-UNN shows again a drastic loss from 0–4 target likes, falling even below POP. In terms of catalog coverage, the trends are very similar to the ones in movies domain. Interestingly, CD-INN beats again all the other methods in terms of accuracy and catalog coverage with 1 and 2 likes in the target domain. Analyzing the use of cross domain information, INN is once again able to reach better accuracy using the additional movie likes, while HeteRec obtains a benefit where less feedback is provided (from 1 to 5). Again, CD-HeteRec with 0 likes in the target domain overcomes all the other methods even where they use more target information (up to 8). However its catalog coverage is too low (1%) compared the other methods (>10%).

Summing up, we may say that cross-domain information is arguably useful to face the cold-start user problem, allowing to generate relevant recommendation even where no target information is available. The choice of the method depends on the domain and amount of user information available. Moreover, we discover that some methods obtain exceptionally better results using only the source domain rather than using a few target feedbacks as well. More research will be needed for better understanding this trend.

Cross-domain recommendation diversity.

This section addresses RQ2, namely testing whether cross-domain information leads to more diverse recommendations. Tables 1 shows the diversity results for movies and music domain in terms of BinomDiv@10. We also compared the methods using the ILD metric, but we do not show its values, since they obtain a very similar trend to BinomDiv.

Music–Movies. POP obtains good results values, since all the most popular movies in the dataset belong to different genres, but CD-UNN and UNN overcome it with 1 and 2 likes, and only UNN from 3 to 6. In general, using cross-domain music information yields to less diverse recommendations. **Movies–Music**. PathRank and CD-PathRank produce the most diverse recommendations. Conversely, MF methods lead to the worst diversity. In contrast to the previous situation, using cross-domain movies information for music recommendations improves nearly always the diversity degree of the recommendations.

Size and diversity of source domain user profiles.

In order to address RQ3, we compute the number of preferences and the intra-list diversity of the user profiles in the source domain, and group users in different ranges. For the profile sizes we split users in intervals of 20 likes, from size 0 to 100 and beyond, and for profile diversity we classify the users in terms of the distribution of ILD scores. Specifically, we define four groups based on the quartiles which we name `Very low` (0–25%), `Low` (25%–50%), `Medium` (50%–75%), and `High` (75%–100%). Finally, we average the MRR of the recommendation lists in the target domain separately for each group, first in terms of profile size and then in terms of diversity. Figure 1 shows the relation between the quality of the target recommendations and the analysed source profile properties. We only report the results for the extreme cold-start profile sizes in the target, i.e., 0 and 10, as the rest showed similar behavior.

In terms of source profile size, we notice that in general the quality of target recommendations improves monotonically as more information about the user's preferences is available in the source domain. This trend holds for all the evaluated algorithms with the exception of CD-IMF in music, where we see that the performance degrades when the size of the source profile is larger than 100. In this case, we argue that the abundance of auxiliary preferences could be drifting the learning of the model parameters towards the source domain, although a deeper analysis is needed to confirm our intuition.

Regarding the impact of the source profile diversity we find that the best results are achieved for users very focused on limited types of items, whereas a more diverse profile has a negative effect on the accuracy of the recommendations.

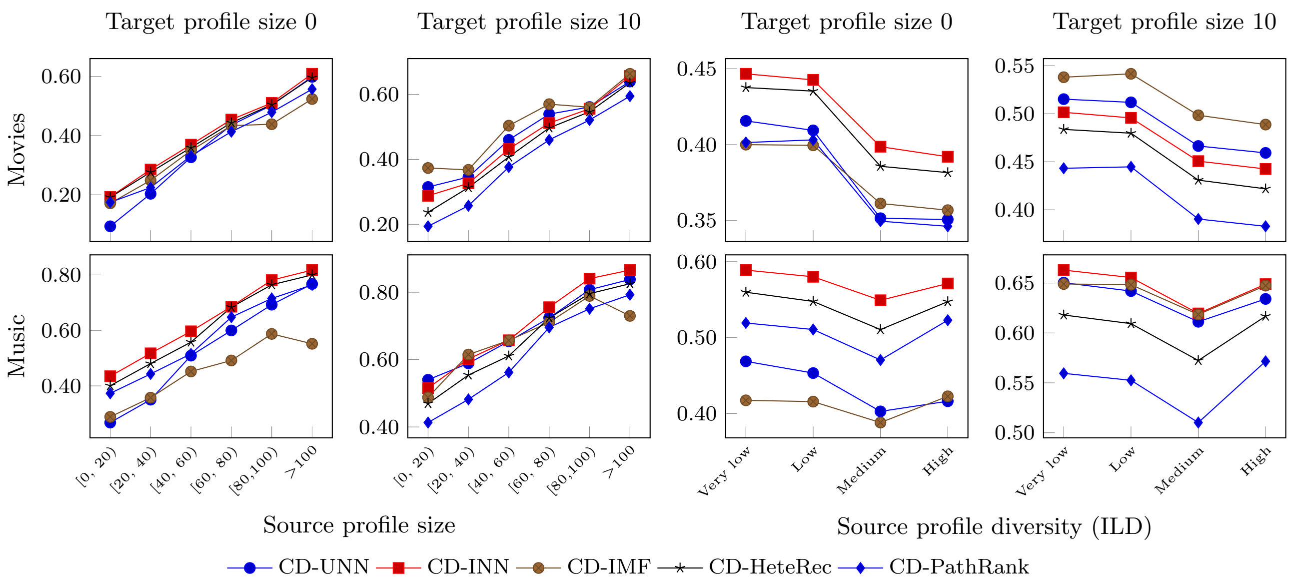

Figure 1: MRR values of the cross-domain recommendation methods for different user profile sizes and profile diversity values in the source domain. Each row corresponds to the two target domains in our dataset.

This seems to indicate that the evaluated algorithms struggle to find inter-domain correlations, specially from music to movies. In the case of very high diversity we see that the two settings diverge: variety in source movie preferences is beneficial for music recommendations, whereas the converse has the opposite effect.

We conclude that both the source user profile size and diversity have a significant impact on the quality of cross domain recommendations, thus confirming RQ3. On a side note, we observe the superior performance of CD-INN in most of the considered scenarios, specially in the extreme cold-start with target profile size of 0. We argue that this behavior is a consequence of the relatively large overlap of users between the analysed domains, an issue that we plan to further investigate in future work.

4. CONCLUSIONS AND FUTURE WORK

We have studied the quality of cross-domain recommendations in terms of accuracy, diversity and catalog coverage, evaluating a number of algorithms on two datasets with positive-only feedback. Our results show the benefits of cross-domain information in cold-start situations in terms of ranking accuracy. Regarding diversity we observe different behavior in the two datasets, and therefore conclude that in general the results depend on the target domain. We have also studied the impact of the size and diversity of user profiles in the source domain, concluding that while more cross-domain user preferences are helpful, a greater item diversity in the source profile can actually harm the performance in the target domain. Following this work we intend to further investigate which characteristics of the datasets could explain the differences we found in both recommendation and user profile diversity. We plan to extend our analysis to more domains, e.g. books, and to evaluate more sophisticated methods from the state of the art, such as [5].

Acknowledgments

This work was supported by the Spanish Ministry of Science and Innovation (TIN2013-47090-C3-2) and by the project PON03 PE_00136_1 Digital Services Ecosystem: DSE.

5. REFERENCES

[1] I. Cantador, I. Fernández-Tobías, S. Berkovsky, and P. Cremonesi. Cross-domain recommender systems. In *Recommender Systems Handbook*, pp. 919–959. Springer, 2015.

[2] T. Di Noia, V. C. Ostuni, J. Rosati, P. Tomeo, and E. Di Sciascio. An analysis of users' propensity toward diversity in recommendations. In *RecSys '14*, pp. 285–288, 2014.

[3] M. Enrich, M. Braunhofer, and F. Ricci. Cold-start management with cross-domain collaborative filtering and tags. In *EC-Web '13*, pp. 101–112, 2013.

[4] I. Fernández-Tobías, M. Braunhofer, M. Elahi, F. Ricci, and I. Cantador. Alleviating the new user problem in collaborative filtering by exploiting personality information. *User Model. User-Adapt. Interact.*, 26(2-3):221–255, 2016.

[5] L. Hu, J. Cao, G. Xu, L. Cao, Z. Gu, and C. Zhu. Personalized recommendation via cross-domain triadic factorization. In *WWW '13*, pp. 595–606, 2013.

[6] Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *ICDM '08*, pp. 263–272, 2008.

[7] D. Kluver and J. A. Konstan. Evaluating recommender behavior for new users. In *RecSys '14*, pp. 121–128, 2014.

[8] S. Lee, S. Park, M. Kahng, and S.-g. Lee. Pathrank: A novel node ranking measure on a heterogeneous graph for recommender systems. In *CIKM '12*, pp. 1637–1641, 2012.

[9] V. C. Ostuni, T. Di Noia, E. Di Sciascio, and R. Mirizzi. Top-n recommendations from implicit feedback leveraging linked open data. In *RecSys '13*, pp. 85–92, 2013.

[10] N. Rubens, M. Elahi, M. Sugiyama, and D. Kaplan. Active learning in recommender systems. In *Recommender Systems Handbook*, pp. 809–846. Springer, 2015.

[11] S. Sahebi and P. Brusilovsky. It takes two to tango: An exploration of domain pairs for cross-domain collaborative filtering. In *RecSys '15*, pp. 131–138, 2015.

[12] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *SIGIR '02*, pp. 253–260, 2002.

[13] P. Tomeo, I. Fernández-Tobías, T. Di Noia, and I. Cantador. Exploiting linked open data in cold-start recommendations with positive-only feedback. In *CERI '16*, 2016.

[14] S. Vargas, L. Baltrunas, A. Karatzoglou, and P. Castells. Coverage, redundancy and size-awareness in genre diversity for recommender systems. In *RecSys '14*, pp. 209–216, 2014.

[15] P. Winoto and T. Y. Tang. If you like the devil wears prada the book, will you also enjoy the devil wears prada the movie? A study of cross-domain recommendations. *New Generation Computing*, 26(3):209–225, 2008.

[16] X. Yu, X. Ren, Y. Sun, Q. Gu, B. Sturt, U. Khandelwal, B. Norick, and J. Han. Personalized entity recommendation: A heterogeneous information network approach. In *WSDM '14*, pp. 283–292, 2014.