Implicit Feedback Recommendation via Implicit-to-Explicit Ordinal Logistic Regression Mapping

Denis Parra
University of Pittsburgh

Alexandros Karatzoglou
Telefonica Research

Xavier Amatriain
Telefonica Research

Idil Yavuz
University of Pittsburgh

ABSTRACT

One common dichotomy faced in recommender systems is that explicit user feedback -in the form of ratings, tags, or user-provided personal information- is scarce, yet the most popular source of information in most state-of-the-art recommendation algorithms, and on the other side, implicit user feedback - such as numbers of clicks, playcounts, or web pages visited in a session- is more frequently available, but there are fewer methods well studied to provide recommendations based on this kind of information.

Given the current scenario, and under a situation where just implicit user feedback is available, it would be more appropriate either to provide recommendations using the implicit data and implicit-fedback-based methods, or to map implicit user feedback to explicit feedback and then use an explicit-based algorithm? On this paper, we analyze this problem in the context of music recommendation by means of a well-known implicit feedback recommendation method described in Hu et al. [1] by comparing the use of raw playcounts with the use of explicit data - user ratings - obtained by mapping implicit to explicit feedback with a novel mixed-effects logistic regression model.

1. INTRODUCTION

Recommender Systems (RS) [2] have proved their business value and impact on many application scenarios that go from recommending movie rentals to new contacts on a social network. One of the main features of these systems is that they rely on understanding user preferences in order to estimate the *utility* of items and decide whether they should be recommended. These user preferences are infered by taking into account direct feedback from the user, either in *explicit* or *implicit* form.

We obtain implicit feedback [3] by measuring the interaction of the user with the different items. We can use signals such as the number of playcounts in a song, or the clicks on webpages as implicit feedback. This kind of data is obtained without incurring into any overhead on the user, since it is obtained from direct usage [4]. However, it is not clear that we can trust a simple one-to-one mapping between usage and preference [5]. On the other hand, explicit feedback is obtained by directly querying the user, who is usually presented with an integer scale where to quantify how much she likes the items. In principle, explicit feedback is a more robust way to extract preference, since the user is reporting directly on this variable, removing the need of an indirect inference. However, it is also known that this kind of feedback is affected by user inconsistencies known as *natural noise* [6]. Besides, the fact that we are introducing a user overhead, makes it difficult to have a complete view on the user preferences [7].

None of the two existing strategies for capturing user feedback clearly outperforms the other. Ideally, we would like to use implicit feedback, minimizing the impact on the user, but having a robust and proven way to map this information to the actual user preference. In a previous work [8], we tested several regression models and we were able to map implicit user feedback to explicit ratings. Our results were satisfactory, but we did not compare to state-of-the-art methods that make use of raw implicit information to provide recommendations. In this paper we propose an ordinal logistic regression model that by using a few ratings is able to infer a generic parametric mapping from implicit to explicit data. Our mapping model integrates usual implicit user feedback (playcounts) with contextual information (how recently the user listened to an album). We compare our approach to a state-of-the art algorithm for implicit feedback recommendations and discuss possible extensions.

2. PRELIMINARIES AND RELATED WORK

Implicit feedback is much more readily available in practical scenarios for recommender systems. However, most of the research literature focuses on the use of explicit feedback input since this is considered the ground truth on the user preferences and allows to reduce the recommender problem to one of predicting ratings.

In one of the few papers addressing the implicit feedback recommendation problem [1], Hu *et al.* deal with the implicit feedback recommendation problem by binarizing it and introducing the idea of *confidence*. In our previous work [8], however, we presented an analysis of implicit and explicit feedback that challenged most of the assumptions stated in [1]. In particular: (1) **There is no negative feedback**. While it is true that you cannot interpret "no implicit feedback" as "negative feedback" – and this is true also for explicit feedback–, implicit data can include negative feed-

CARS-2011, October 23, 2011, Chicago, Illinois, USA.

back. You can assume that *low* feedback is negative feedback as long as the granularity of the items is comparable, and there is enough variability. (2) **Implicit feedback is noisy**. Implicit feedback is noisy but, as we showed in previous work [6], so is explicit feedback. (3) **Preference vs. Confidence**. As we showed in our work [8], the numerical value of implicit feedback can indeed be directly mapped to preference, given the appropriate mapping. (4) **Evaluation of implicit feedback**. On the other hand, we do agree that there is no appropriate evaluation approaches for implicit feedback and this is in fact one of the motivations of our work: if we find an appropriate way to map implicit to explicit feedback we can ensure an evaluation that is as good as the one we have in the explicit case.

Our hypothesis that there is some observable correlation between implicit and explicit feedback can be tracked in the literature. Already in 1994, Morita and Shinoda [9] proved that there was a correlation between reading time on online news and self-reported preference. Konstan *et al.* [10] did a similar experiment with the larger user base of the Grouplens project and again found this to be true. Oard and Kim [11] performed experiments using not only reading time but also other actions like printing an article to find a positive correlation between implicit feedback and ratings. Koh *et al.* did a thorough study of rating behavior in two popular websites [12]. They hypothesize that the overall popularity or average rating of an item will influence raters and they conclude that while there is an effect, this depends on the cultural background of the raters.

Lee *et al.* [13] implement a recommender system based on implicit feedback by constructing "pseudo-ratings" using temporal information. In this work, the authors introduce the idea that recent implicit feedback should contribute more positively towards inferring the rating. The authors also use the idea of distinguishing three temporal bins: old, middle, and recent.

Two recent works approach the issue of implicit feedback in the music domain. Jawasher *et. al* analyze the characteristics of user implicit and explicit feedback in the context of last.fm music service [14]. However, their results are not conclusive due to limitations in the dataset since they only used explicit feedback available in the last.fm profiles, which is limitted to the *love/ban* binary categories. This data is very sparse and, as the authors report, almost non-existant for some users or artists. On the other hand, Kurdomova *et. al* use a Bayesian approach to learn a classifier on multiple implicit feedback variables [15]. Using these features, the authors are able to classify liked and disliked items with an accuracy of 0.75, uncovering the potential of mapping implicit feedback directly to preferences.

In our previous work [8], we showed that it was possible to create a simple parametric model for implicit feedback by using linear regression on some available explicit ratings. However, as we will explain, in the context of user ratings, it may be more appropriate to use a mixed-effects ordinal logistic regression model. In this context, the main contribution of our present work is to present an ordinal logistic regression model that allows to map implicit data into explicit ratings for the task of recommendation. We make our model context-aware with respect to how recently a user listened to an album by *contextual modeling*, i.e., using the contextual information directly in the modelling technique, unlike data-driven approaches such as contextual *pre-filtering* or *post-filtering* [16]. Once the implicit-to-explicit mapping is performed, we can use the inferred ratings in methods for explicit or implicit data. We can then compare the performance of these models to the one by Hu *et al.* in several experiments.

3. REGRESSION MODELS

3.1 Linear Regression

In [8] we introduce a linear regression model to predict explicit preference of users on music albums in the form of ratings based on implicit user behavior variables - (1) *Implicit Feedback (if)*: playcount for a user on a given item; (2) *Global Popularity (gp)*: global playcount for all users on a given item; (3) *Recentness (re)* : time elapsed since user played a given item. In that article, we compare different linear regression models based on the aforementioned variables and we find that the variables implicit feedback and recentness explain the largest part the variability of the ratings, while global popularity explained a very small portion. This result suggested us that the two former variables would be better predictors of the user preference, and we supported these assumption by performing a 10-fold cross validation experiment using the data of our online survey on music preference as a ground truth. The RMSE values were consistent with the previously described regression analysis.

3.1.1 Limitations and shortcomings of Linear Regression

Although the linear regression gives good results, there are some considerations that must be observed to generalize this model to other domains and to make it able to be compared with other approaches. First, depending on the application we may want the predicted values to fall in the range from 1 to 5, but using linear regression we cannot ensure it. Second, as in most of recommender systems research, our main evaluation metric is RMSE. When using this metric, we are assuming that ratings form an interval scale, i.e. the distance between any two consecutive values in the rating scale is the same. However, in a previous study [6], we have shown that users have a larger probabilty to be more inconsistent with some ratings numbers than with others, what give us the clue that users do not see the rating scale as equally spaced. Hence, we should consider the ratings as an ordinal variable rather than an linear or interval one. This also implies that RMSE is not a good measure alone to predict user preference, it should be combined, and in some cases replaced, with other measures coming from Information Retrieval such as precision, recall, or nDCG.

Given that users present individual variability in their ratings, a good extension of our model should include the user as a random factor. Additionally, given that ratings are actually an ordinal variable, as explained in the previous paragraph, and the fact that are not normally distributed, logistic regression is a proper alternative to our linear regression model. Combining both considerations, our next model for implicit-to-explicit behavior mapping model will be a mixed-effects logistic regression.

3.2 Mixed-effects Ordinal Logistic Regression

The multinomial logistic regression is the natural model for an ordinal scale variable (rating, that ranges from 1 to 5) and a mixed-effects model will help us to reduce the vari-

Effect	Estimate	*SE*	*DF*	*t*	Pr > |*t*|
intercept 1	−1.2740	0.2808	112	−4.54	<.0001
intercept 2	0.3791	0.2784	112	1.36	0.1759
intercept 3	2.0898	0.2792	112	7.49	<.0001
intercept 4	3.7355	0.2808	112	13.30	<.0001
gp	−0.01589	0.05598	10000	−0.28	0.7766
if	−0.5894	0.08094	10000	−7.28	<.0001
re	−0.04137	0.05395	10000	−0.77	0.4432
gp*if	−0.06955	0.02956	10000	−2.35	0.0187
if*re	−0.1331	0.02782	10000	−4.78	<.0001
concerts	−0.1912	0.07825	10000	−2.44	0.0145

Table 1: Details of the mixed-effects multinomial regression model with 4 fixed effects

ability due to differences in rating among the users. Our multinomial logistic regression, that uses cummulative logit as link function, can be represented as:

$$logit(P(r_{ui} \leq k)) = \alpha_k + X\beta + g_u \quad (1)$$

where $k = \{1, 2, 3, 4\}$, r_{ui} is the rating that user u gives to item i, $P(r_{ui} \leq k)$ is the probability that the rating r_{ui} is less or equal than k, α_k is the intercept for the cumulative probability that rating is less than or equal to k, X is a vector with the actual values of the fixed factors (if, re and gp), β is the vector of coefficients of the fixed factors, $g_u \overset{\text{iid}}{\sim} N(0, \sigma_g^2)$ is the random effect of the users, and

$$logit(p) = \log(\frac{p}{1-p}) \quad (2)$$

To obtain the predicted rating of a user u on an item i, we calculate the expected value of the rating as

$$E[r_{ui}] = \sum_{k=1}^{5} k \cdot P(r_{ui} = k) \quad (3)$$

where

$$P(r_{ui} = k) = \begin{cases} P(r_{ui} \leq k) & , k = 1 \\ P(r_{ui} \leq k) - P(r_{ui} \leq k-1) & , 1 < k < 5 \\ 1 - P(r_{ui} \leq k-1) & , k = 5 \end{cases} \quad (4)$$

4. EXPERIMENTAL SETUP

4.1 Data sets

We use two datasets in this study. The first one was collected by an online user study among users of the last.fm music service between September and October of 2010, containing implicit and explicit information, and also demographic and consumption data. The second one was collected using the last.fm API during May of 2011, and contains only implicit information. The characteristics of both datasets are described in Table 2 .

4.1.1 Generating Explicit Fedback

We conducted an online user study among users of the last.fm music service. The goal of the study was to gather explicit feedback on music albums to compare to the user implicit feedback we obtained by directly crawling the last.fm page related to the user taking the survey. Explicit feedback was obtained by asking users to rate albums on a 1 to 5 star scale. The items to rate were obtained from the list of albums in the user's playlist so that users responded to a personalized survey. Details of this study, such as the strategy to sample the items that were rated by users and the results of user demographics and user consumption, can be found in our previous article [8].

4.1.2 Implicit Music Consumption Feedback

We call *Implicit Music Consumption Feedback* to our *Dataset2* since, unlike Dataset1 that has demographic data of each user, it only has information about implicit behavior of the users: playcount of albums per each user, how recently each album was listened to for the last time, and the total number of listeners of each album in the whole last.fm website. The statistics of this dataset are described in Table 2.

4.2 Regression Model Selection

To select the fixed effects that would be part of our model we conducted a forward selection on the set of all the main effects and their two-way interactions. The main effects considered were if, re, gp (as described in section 3.1) plus ten demographic and consumption variables: gender, age, hours of music per week, hours of internet per week, buying physical records, buying online records, interaction style (preference on listening to tracks or albums), number of concerts per year, interest on reading specialized music blogs or magazines, and familiarity rating music online. We have to pick two models finally because of the nature of our two datasets. In the smallest one (dataset1) we have all the variables obtained by a user study, but in the second dataset (dataset2) we just have implicit information (playcounts per user, how recently the user listened to each album, and the total number of listeners of an album in the whole dataset) that can be reduced to if, re and gp.

After conducting the process of forward selection, the model obtained for dataset1 considers four fixed effects (if, re, gp and concerts per year) and the random effect of the user. The details of the model are described in Table 1. Although the main effects of global popularity (gp) and recentness (re) are not significant, we keep them in the model because their interaction with implicit feedback (if) is significant [17].

For dataset2, we consider in the model if, re, and gp as fixed effects plus the random effect of the user. For the sake of space we do not show the details of this model, but the coefficient and significant values are similar to those shown in Table 1 excepting that the factor *number of concerts* is not considered in the model. As in the previous model, we keep in the model gp and re although they are not significant due to their interaction with if. Under this model, is also

	Dataset1 (Implicit Explicit)	Dataset2 (Implicit)
users	114	2549
albums	6037	6037
entries	10122	111815
density	1.47%	0.73%
avg albums/user	88.79	43.87
avg user/album	1.71	18.52

Table 2: Details of the datasets

	MAP (D1)	nDCG(D1)	MAP(D2)	nDCG(D2)
HK	0.02315	0.14831	0.1014	0.2718
HKlog	0.02742	0.15447	0.1234	0.2954
logit3	0.02636	0.15319	0.1223	0.2944
logit4	0.02601	0.15268	N/A	N/A
popularity	0.48331	0.54378	0.0178	0.1367

Table 3: Results of MAP and nDCG after 5-fold Cross validation on dataset 1 (D1) and dataset 2 (D2)

not significant the intercept for rating equal to 2, which tell us that this intercept is not significantly different than 0, and we may dismiss it from the model.

4.3 Comparing the different approaches

After we have done the implicit-to-explicit mapping, we are in condition to compare the use of impplicit data with inferred explicit data. In this article, we compare four approaches using dataset 1 and three aproaches using dataset 2. The methods we compare, as identified in the first column of Table 3, are:

- *HK*: the implicit feedback method introduced in Hu *et al.* [1] which uses raw playcounts,
- *HKlog*: a variation of the *HK* method, also introduced in [1], that makes a log-transformation of the playcounts,
- *logit3*: the *HK* method, where the input values are the ratings inferred by logistic regression using 3 fixed factors (if, gp, and re)
- *logit4*: similar to *logit3* but adding the factor *number of concerts* in the logistic regression model to infer the ratings. We have this information available just for dataset1.

Description of the HK method. For the implicit feedback modeling we use the Matrix Factorization method developed in [1]. In this Matrix Factorization method a weighted least squares error loss function is minimized. To this end *user-item* interactions p_{ij} are signaled with a 1 and missing interactions are marked with a 0. The counts of *user-item* interactions (e.g. playcounts Y_{ij}) are translated into a confidence measure w_{ij}, which in the case of the *HK* method correspond to $p_{ij} + \alpha Y_{ij}$, and in the case of the *HKlog* method a simple log transform is used where:

$$w_{ij} = \begin{cases} \alpha log(1 + Y_{ijk}) & Y_{ijk} > 0 \\ 1 & Y_{ijk} = 0 \end{cases} \quad (5)$$

This "confidence" is then used as a weight in the loss function and the objective function then becomes

$$\min_{U,M,C} \sum_{i}^{n} \sum_{j}^{m} [w_{ij} \left(p_{ij} - \langle U_{i*} M_{j*} \rangle\right)^2 \quad (6)$$
$$+ \frac{\lambda}{n} ||U_{i*}||^2 + \frac{\lambda}{m} ||M_{j*}||^2]$$

where the Frobenius norm of the factor matrices is used for regularization. This minimization problem is then solved in linear time using Alternate Least Squares and utilizing a trick to avoid direct optimization over the 0 entries of the matrix.

4.3.1 Error Measures

RMSE [18] is probably the most common measure to evaluate the performance of recommender systems and we used it to evaluate and compare our linear regression approaches in [8]. However, when there are no ratings to assess the performance of the algorithms we can not use metrics like RMSE or MAE. Hence, we opt for using Mean Average Precision (MAP) [19] and normalized Discounted Cummulative Gain (nDCG) [20]. The former gives us an overall sense of how well we identify relevant items to recommend from a set of retrieved recommendations, and the latter how well we rank them in a list.

5. RESULTS

In order to evaluate and compare the methods, we split each dataset into 5 groups in order to perform a 5-fold cross validation. The result of each run is a list of recommended items (albums) for each user in the test set, sorted by the preference that the user would have for that item. We calculate MAP and nDCG for each list recommended to a user, judging an item as relevant whether it was consumed (played) at least once by the user. Results can be seen in Table 3.

In the case of dataset 1, the best results of MAP and nDCG are obtained by recommending the most popular items. This result is somewhat expected due to the sparsity of the dataset that affects the methods based on matrix factorization. As shown in Table 2, each album was rated in average by just 1.71 users. This situation is not repeated in dataset 2, where the average number of users per album is 18.52, and then the popularity method performs the worst.

We highlight two results on these initial experiments. The first one is that the log transformation of raw playcounts makes *HKlog* improve clearly over *HK* on both MAP and nDCG measures. The second result we higlight is that *logit3* and *logit4* perform better than *HK* and there is not a big

difference in performance with *HKlog*, leading us investigate further to confirm this difference.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we continue the work that we started in [8] to create a model that allows us to map implicit to explicit user behavior. Using MAP and nDCG metrics, we show that our method is comparable to state of the art methods that provides recommendations making use of implicit user feedback.

The results that we have obtained, part of which we show on this paper, give us some insights but they mainly open research questions that we need to analyze further. We have confirmed in our dataset the benefits of applying a log transformation to the raw user feedback in the Hu *et al.* model, showing consistently better results than the unmodified version.

In terms of the questions we need to further analyze, up to this point, we have considered the factors *implicit feedback* and *global popularity* in our logistic regression models as ordinal variables. We coded these variables on this way to make sure that we were doing an appropiate diverse sampling when creating the user survey described in [8]. However, there is no constraint to rather use the raw playcounts for both factors aforementioned, and we think that this modification can benefit the results of our implicit-to-explicit logistic regression model.

On the experiments run on this study, since we are not predicting user ratings but rather user preference, metrics such as RMSE or MAE can not be used to compare the methods so we opt for IR metrics such as MAP and nDCG, which rely on how we define *relevancy*. We wonder if our definition of relevance might bias our results and conclusions. As we have stated it before, we think that low feedback might be, in fact, negative feedback. For this reason, we are currently testing different user activity (implicit feedback) thresholds to define relevancy in order to analyze how that influences the evaluation of the different recommendation approaches.

7. REFERENCES

[1] Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *Proceedings of ICDM 2008*, 2008.

[2] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011.

[3] Douglas Oard and Jinmook Kim. Implicit feedback for recommender systems. In *in Proceedings of the AAAI Workshop on Recommender Systems*, pages 81–83, 1998.

[4] G. Potter. Putting the collaborator back into collaborative filtering. In *2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, 2008.

[5] D. M. Nichols. Implicit rating and filtering. In *In Proceedings of the Fifth DELOS Workshop on Filtering and Collaborative Filtering*, pages 31–36, 1997.

[6] X. Amatriain, J.M. Pujol, and N. Oliver. I like it... i like it not: Evaluating user ratings noise in recommender systems. In *Proc. of the 2009 Conference on User Modeling, Adaptation, and Personalization*, 2009.

[7] G. Jawaheer, M. Szomszor, and P. Kostkova. Characterisation of explicit feedback in an online music recommendation service. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 317–320, 2010.

[8] D. Parra and X. Amatriain. Walk the talk: Analyzing the relation between implicit and explicit feedback for preference elicitation. In *Proc. of the 2011 Conference on User Modeling, Adaptation, and Personalization*, 2011.

[9] M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference*, pages 272–281, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[10] Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl. Grouplens: applying collaborative filtering to usenet news. *Commun. ACM*, 40(3):77–87, 1997.

[11] D. Oard and J. Kim. Modeling information content using observable behavior. In *Proc. of the ASIST Annual Meeting*, pages p481–88, 2001.

[12] N.S. Koh, N. Hu, and E. K. Clemons. Do online reviews reflect a product's true perceived quality? - an investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications*, 2010.

[13] T. Lee, Y. Park, and Y. Park. A time-based approach to effective recommender systems using implicit feedback. *Expert Syst. Appl.*, 34(4):3055–3062, 2008.

[14] Gawesh Jawaheer, Martin Szomszor, and Patty Kostkova. Comparison of implicit and explicit feedback from an online music recommendation service. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, 2010.

[15] S. Kordumova, I. Kostadinovska, M. Barbieri, V. Pronk, and J. Korst. Personalized implicit learning in a music recommender system. In *UMAP 2010*, 2010.

[16] Gediminas Adomavicius and Alexander Tuzhilin. Context-aware recommender systems. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 217–253. Springer US, 2011.

[17] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*. Irwin, Chicago, 1996.

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

[19] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[20] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20:422–446, October 2002.