Beyond Rating Prediction Accuracy: On New Perspectives in Recommender Systems

Panagiotis Adamopoulos
Department of Information, Operations and Management Sciences
Leonard N. Stern School of Business, New York University
padamopo@stern.nyu.edu

ABSTRACT

This paper proposes a number of studies in order to move recommender systems beyond the traditional paradigm and the classical perspective of rating prediction accuracy. We contribute to existing helpful but less explored paradigms and also propose new approaches aiming at more useful recommendations for both users and businesses. Working toward this direction, we discuss the studies we have conducted so far and present our future research plans. In particular, we move our focus from even more accurate rating predictions and aim at offering a holistic experience to the users by avoiding the over-specialization of generated recommendations and providing the users with sets of non-obvious but high quality recommendations that fairly match their interests and they will remarkably like.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems - Human Factors; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - Information filtering, Selection process; H.4.m [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Design, Experimentation, Human Factors, Measurement, Performance, Theory

Keywords

Item Accuracy; Unexpectedness; Novelty; Serendipity; Diversity; Recommendation Opportunities; Recommendation Sets; Recommender Systems

1. INTRODUCTION

Over the last two decades, a wide variety of different types of recommender systems (RSs) has been developed and successfully used across several domains [6]. During this time, many researchers have focused mainly on the development and improvement of efficient algorithms for more accurate rating prediction. Although the recommendations of the latest class of systems are significantly more accurate than they used to be a decade ago [7] and the broad social and business acceptance of RSs has already been achieved, there is still a long way to go in terms of satisfaction of users' actual needs [14]. This is due, primarily, to the fact that many existing RSs focus on providing more accurate rather than more useful recommendations. Some of the main problems pertaining to this narrow rating prediction accuracy-based focus of many existing RSs [8] and the ways to broaden the current approaches have been discussed in [18].

Even though the aforementioned rating prediction perspective is the prevailing paradigm in recommender systems, there are other perspectives that try to alleviate the problems pertaining to this narrow rating prediction accuracy-based focus and have been gaining significant attention in the field of RS [10]. In particular, some of the most recent perspectives maintain that RSs should make the users familiar with the various product categories and the whole product catalog. In addition, recommender systems should provide personalized recommendations from a wide range of items and should also enable the users to find relevant items that might be hard to discover. Also, they should increase user satisfaction and engagement and offer a superior user experience. Moreover, RSs should also be able to reduce user search costs, improve the quality of decisions that consumers make, and increase their welfare. Besides, from a business perspective, RSs should increase the sales volume and conversion rates, as well as promote items from the long tail that usually exhibit significantly lower marginal cost and, at the same time, higher marginal profit.

Moving beyond the classical perspective of the rating prediction accuracy, the main objective of this stream of research is to contribute to existing helpful but less explored paradigms of RSs as well as to propose new approaches that will result in more useful recommendations for both users and businesses. Working toward this direction, we discuss the studies we have conducted so far and present our future research plans. In particular, we move our focus from even more accurate rating predictions and aim at offering a holistic experience to the users by avoiding the over-specialization of recommendations and providing the users with non-obvious but high quality recommendation sets (not restricted to simple lists of individual items) that fairly match their interests and they will remarkably like.

Other streams of research that improve recommender systems going beyond rating prediction accuracy include work on *human-computer interaction* (HCI), which involves the

RecSys'13, October 12–16, 2013, Hong Kong, China.
ACM 978-1-4503-2409-0/13/09.
http://dx.doi.org/10.1145/2507157.2508073.

study and design of the interaction between users and RSs, *diversification*, which maximizes the variety of items in a recommendation list, *group recommenders* [19], which recommend items for groups of people, rather than individuals, and recommendation *sequences* [17], where sequences of ordered items are recommended instead of single items.

2. RESEARCH STUDIES

In this stream of research, we are moving beyond the classical perspective of the rating prediction accuracy and we are aiming at providing even more useful recommendations for both users and businesses. Working toward this direction, we designed a number of research studies to explore issues related to several new approaches for providing recommendations. In particular, Section 2.1 proposes a concept of *unexpectedness* and specific metrics to measure both unexpectedness and quality of recommendations, as well as algorithms for generating such recommendations. Then, Section 2.2 introduces a method for generating recommendations based on *recommendation opportunities* using higher *weighted percentiles*. Finally, Section 2.3 discusses *recommendation sets* where the generated recommendation lists are based on sets of candidate items, rather than just individual items, also taking into account various interactions among items, such as complementarity and substitution effects. All these research studies go beyond the classical perspective of RSs and aim at providing the users with non-obvious but high quality recommendation sets that fairly match their interests and they will remarkably like. Thorough discussions of these concepts, implementation details, and experimental results are provided in [2, 5, 4, 3].

2.1 Beyond over-specialization

Sections 2.1.1 and 2.1.2 address the over-specialization problem in recommender systems focusing on *unexpectedness* and the related concepts of coverage, novelty, serendipity, and diversity of recommendation lists.

2.1.1 Expecting the Unexpected

In [2, 3], we propose a concept of unexpected recommendations as recommending those items that significantly depart from the expectations of the users and suggest a method for generating such recommendations, based on the utility theory of economics, as well as specific metrics to measure the unexpectedness of recommendation lists.

In particular, we formally define the concept of *unexpectedness* in recommender systems taking into account the actual *expectations* of the users and discuss how the concept of unexpectedness is differentiated from various related notions, such as novelty, serendipity, and diversity. Following the Greek philosopher Heraclitus, we approach this difficult problem of finding and recommending unexpected items by first capturing the items expected by the users. Toward this direction, we suggest several mechanisms for specifying users' expectations that can be applied across various domains. Such mechanisms include the past transactions performed by the users, knowledge discovery and data mining techniques, and experts' domain knowledge. Besides, we formulate and fully operationalize the notion of unexpectedness and present an algorithm for providing unexpected recommendations of high quality that are hard to discover but fairly match the users' interests, based on the *utility theory* of economics. Moreover, we propose specific performance metrics to measure the unexpectedness of the generated recommendation lists taking into account also the usefulness of individual items.

Using "real-world" data sets, various examples of sets of expected recommendations, and different utility functions and distance metrics, we were able to test the proposed method under a large number of experimental settings including various levels of sparsity, different mechanisms for specifying users' expectations, and different cardinalities of these sets of expectations. The empirical study showed that all the examined variations of the proposed method significantly outperformed in terms of unexpectedness the standard baseline algorithms, including item-based and user-based k-Nearest Neighbors [13], Slope One [16], and Matrix Factorization [15]. This demonstrates that the proposed method indeed effectively captures the concept of unexpectedness since, in principle, it should do better than unexpectedness-agnostic methods. Furthermore, the proposed method for unexpected recommendations performed at least as well as, and in some cases even better than, the baseline algorithms in terms of the classical accuracy-based measures, such as root-mean-square error (RMSE) and the F-measure, as well as other popular performance measures, such as catalog coverage, aggregate diversity, serendipity, and the Gini coefficient. In addition, we presented a number of actual recommendation examples generated by the proposed method and the employed baseline approaches and provided insightful qualitative comments.

One of the main premises of the proposed method is that the users' expectations should be explicitly considered in order to provide the users with unexpected recommendations of high quality that are hard to discover but fairly match their interests. Hence, the greatest improvements both in terms of unexpectedness and accuracy vis-à-vis all other approaches were observed in the experiments using the more accurate sets of expectations. Moreover, the use of a utility function of standard form illustrates that the proposed method can be easily implemented in existing recommender systems as a new component that enhances unexpectedness of recommendations, without the need to further modify the current rating prediction procedures.

2.1.2 Probabilistic Neighborhoods

In [4], we propose a new probabilistic method for neighborhood selection in collaborative filtering (CF) models. In particular, we illustrate the practical implementation of the proposed approach presenting a specific variation of the classical k-nearest neighbors (k-NN) CF method in which the neighborhood selection is based on an underlying probability distribution instead of just the k neighbors with the highest similarity level to the target user. For the probabilistic neighborhood selection, we use an efficient method for weighted sampling [21] of k neighbors without replacement that also takes into consideration the similarity levels between the target user and the n candidate neighbors. The key intuition for this *probabilistic nearest neighbors* collaborative filtering method is two-fold. First, using the neighborhood with the most similar users to estimate unknown ratings and recommend candidate items, the generated recommendation lists usually consist of known items with which the users are already familiar. Second, because of the multidimensionality of users' tastes, there are many items that the target user may like and are unknown to the k most similar users to her/him. Thus, we propose the use of probabilistic neighborhood selection in order to alleviate the aforementioned problems and move beyond the limited focus of rating prediction accuracy.

To investigate this claim, we conducted an empirical study and we tested the proposed method under a large number of experimental settings. In detail, we used a large number of probability distributions from different families of distributions with various location and shape parameters, in order to compare the proposed probabilistic method for neighborhood selection against the standard collaborative filtering approach in terms of popular evaluation metrics for item prediction accuracy, utility-based ranking, coverage, diversity, unexpectedness, and dispersion of recommendations.

The experimental results illustrate that the proposed approach indeed generates recommendations that are orthogonal to the classical CF method. We also demonstrated that the proposed method performs at least as well as, and in some cases even better than, the standard user-based CF approach in terms of popular item prediction accuracy and utility-based ranking measures, such as the F-measure and the normalized cumulative discounted gain, across various experimental settings. Besides, we showed that the performance improvement is not achieved at the expense of some other popular performance measures that go beyond the rating prediction accuracy, such as catalog coverage, aggregate diversity, and recommendation dispersion and mobility.

2.2 Recommendation Opportunities

In [5], under a definition of a *recommendation opportunity* as how much a user *could* realistically like an item, we aim at recommending items that the users will remarkably like. Moving beyond the standard perspective of rating prediction accuracy and exploring such recommendation opportunities can increase user satisfaction and engagement and offer a superior user experience through the discovery of items that the users will really like.

In particular, we illustrate the practical implementation of the proposed approach presenting a certain variation of the classical user-based k-NN collaborative filtering method in which the estimation of an unknown rating of the target user for an item is based not on the weighted averages of the k nearest neighbors but on the *weighted percentile* of the ratings of these k neighbors. For the estimation of the weighted percentile of the distribution of the ratings in the neighborhood of the target user, an efficient method is used that does not increase the computational complexity of the classical k-NN method. The key intuition behind the weighted percentile method, instead of using weighted averages, is that high percentiles (such as in the 70% to 90% range) constitute more realistic estimates of how much a targeted user *could* possibly like the candidate item. As a consequence, the proposed approach not only provides more useful for the users recommendations of items that they remarkably like but also has the potential to let us better identify and serve any specific niches of the market.

To support this claim, we conducted an empirical study and showed that the proposed percentile method outperforms by a wide margin, across various experimental settings, the standard user-based CF approach in terms of item prediction accuracy measures, such as precision, recall, and the F-measure, and also utility-based ranking metrics, such as normalized cumulative discounted gain and mean reciprocal rank. Finally, we demonstrated that this performance improvement is not achieved at the expense of some other popular performance measures, such as catalog coverage, aggregate diversity, and the Gini coefficient. This illustrates that our proposed weighted percentile method for *recommendation opportunities* performs at least as well as, or even significantly better than, the classical user-based collaborative filtering method in terms of these important measures, in most of the experiments.

Nevertheless, apart from the user-based and item-based k-NN CF approaches, other popular RSs methods that can be easily extended, with the use of quantile regression [12], in order to allow us both to build models that predict high percentiles and to evaluate them with regard to the goal of predicting percentiles of estimated ratings, include content-based methods and Matrix Factorization [11].

2.3 Recommendation Sets

In the classical recommendation system paradigms, the generated recommendation lists are based on the top-N items with the highest estimated ratings regardless of the possible interaction effects among the candidate items. One exception of this paradigm is the concept of diversity, according to which the variety of items in a recommendation list is maximized. However, in many recommendation settings there are important interactions among the candidate items that should be explicitly considered. For instance, in a clothing on-line retail setting the utility of recommending to a user a specific pair of pants depends on whether a matching shirt is also included in the same recommendation list or not. Similarly, in the case of a RS for a supermarket, the utility of recommending olive oil to a user might depend on whether feta cheese and tomatoes are recommended as well.

In our next study, moving beyond the classical recommendation lists of individual items with the highest estimated ratings and working toward our objective of providing the users with non-obvious but high quality recommendation sets that fairly match their interests and they will really like, we will propose to generate *recommendation sets*, rather than individual items, taking into account various interactions effects among the candidate items [9], the potential prerequisites and constraints of the items [20], and the limited budget of the users [22]. In particular, we will focus on co-occurrence interaction effects, such as complementarity and substitution effects, aiming at developing effective methods for both accurately estimating the various effects and efficiently recommending such sets of items.

To experimentally test the proposed approach we plan to conduct a live experiment involving human subjects. In detail, we plan to design an on-line experiment and examine the proposed paradigm in the context of massive open online courses (MOOCs) [1]. As part of the experiment, we will generate personalized recommendations for MOOCs taking into account the academic major of the human subject, the various dependencies among the courses, such as the complementarity and substitution effects, and different prerequisites and constraints, such the difficulty and the estimated workload of each recommended course.

3. DISCUSSION AND FUTURE WORK

Successfully completing the aforementioned work will help the recommender systems field move further beyond the perspective of rating prediction accuracy. Following the proposed stream of research and adhering to our main objective of providing more useful recommendations for both users and businesses, we both contribute to existing helpful but less explored paradigms for recommender systems and propose new valuable approaches and perspectives. Working toward this direction of providing more useful recommendations for both users and businesses, we discussed the studies we have conducted so far and also presented in detail some of our future research plans. In particular, the studies discussed in Section 2 move our focus from even more accurate rating predictions and aim at offering a holistic experience to the users by avoiding the over-specialization of recommendations and providing the users with non-obvious but high quality recommendation sets that fairly match their interests and they will remarkably like.

In detail, Section 2.1 proposes a concept of *unexpectedness* and specific metrics to measure both unexpectedness and quality of recommendations, as well as algorithms for generating such recommendations, Section 2.2 introduces a method for generating recommendations that the users will remarkably like based on *recommendation opportunities* using higher *weighted percentiles*, and Section 2.3 discusses *recommendation sets*, rather than individual items, taking into account various interactions among items, such as complementarity and substitution effects.

As part of the future work, we would like to implement and evaluate the proposed approaches in a traditional on-line retail setting as well as in a platform for massive open on-line courses (MOOCs) [1]. Besides, we would like to conduct a series of live controlled experiments with human subjects in order to study the on-line user behavior, examine and actively adjust the trade-off between exploration (e.g. unexpectedness, serendipity, diversity, etc.) and exploitation (e.g. accuracy) of recommender systems, and better evaluate the proposed perspectives in a user-centric framework. In addition, we plan to investigate how the proposed recommendation approaches and perspectives can be effectively combined with traditional approaches in hybrid recommender systems aiming at accurate and non-obvious recommendation sets that the users will remarkably like.

Furthermore, the generated recommendations should be useful for both users and businesses and, hence, as part of our future work, we would also like to focus more on the business perspective and the business value aspects of recommender systems. Thus, we would like to design an empirical study in order to examine the economic impact of the proposed recommender system approaches and perspectives across various domains and recommendation settings. In particular, we would like to study and estimate the economic impact of both the direct effect of offering recommendations from a wider range of items that are harder to discover and the indirect effect of recommending items from the long tail and not focusing mostly on bestsellers that usually exhibit higher marginal cost. Such a research study can shed more light on the usefulness of recommender systems for businesses and further promote the use of non-classical perspectives and approaches that go beyond the traditional paradigm of rating prediction accuracy.

Moreover, the proposed program of research can be extended in various other valuable directions. For instance, the main strengths and weaknesses of the different existing recommender system algorithms, in terms of the multiple objectives of the aforementioned perspectives and metrics, should be thoroughly examined and, if possible, the "best" mixture in each context among all the existing and proposed approaches should be identified. Finally, the possible spillover effects of the proposed approaches on other important objectives and properties of recommender systems, such as the trustworthiness and persuasiveness of recommendations, should be extensively examined as well.

4. ACKNOWLEDGMENTS

I would like to thank my advisor and co-author, Prof. Alexander Tuzhilin, for his support, guidance, and valuable feedback.

5. REFERENCES

[1] P. Adamopoulos. What Makes a Great MOOC? An Interdisciplinary Analysis of Student Retention in Online Courses. In *Proceedings of the 34th International Conference on Information Systems*, ICIS '13, 2013.

[2] P. Adamopoulos and A. Tuzhilin. On Unexpectedness in Recommender Systems: Or How to Expect the Unexpected. In *DiveRS 2011*, RecSys '11. ACM, 2011.

[3] P. Adamopoulos and A. Tuzhilin. On Unexpectedness in Recommender Systems: Or How to Better Expect the Unexpected. *Working Paper: CBA-13-03, New York University*, 2013. http://ssrn.com/abstract=2282999.

[4] P. Adamopoulos and A. Tuzhilin. Probabilistic Neighborhood Selection in Collaborative Filtering Systems. *Working Paper: CBA-13-04, New York University*, 2013. http://hdl.handle.net/2451/31988.

[5] P. Adamopoulos and A. Tuzhilin. Recommendation Opportunities: Improving Item Prediction Using Weighted Percentile Methods in Collaborative Filtering Systems. In *Proceedings of RecSys '13*. ACM, 2013.

[6] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6):734–749, June 2005.

[7] R. M. Bell, J. Bennett, Y. Koren, and C. Volinsky. The million dollar programming prize. *IEEE Spectr.*, 46(5):28–33, 2009.

[8] P. Cremonesi, F. Garzotto, S. Negro, A. V. Papadopoulos, and R. Turrin. Looking for "good" recommendations: a comparative evaluation of recommender systems. In *INTERACT '11*, 2011.

[9] D. L. Hansen and J. Golbeck. Mixing it up: recommending collections of items. In *Proceedings of CHI '09*. ACM, 2009.

[10] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.

[11] A. Karatzoglou and M. Weimer. Quantile matrix factorization for collaborative filtering. *In Proceedings of EC-Web '10*, 2010.

[12] R. Koenker. *Quantile Regresssion*. Wiley Online Library, 2005.

[13] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. Grouplens: applying collaborative filtering to usenet news. *Comm. of the ACM*, 40(3), 1997.

[14] J. A. Konstan and J. T. Riedl. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22:101–123, 2012.

[15] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8), 2009.

[16] D. Lemire and A. Maclachlan. Slope one predictors for online rating-based collaborative filtering. *CoRR*, abs/cs/0702144, 2007.

[17] J. Masthoff. Group recommender systems: Combining individual models. In *Recommender Systems Handbook*, pages 677–702. Springer, 2011.

[18] S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *Proceedings of CHI '06*, pages 1097–1101. ACM, 2006.

[19] M. O'Connor, D. Cosley, J. A. Konstan, and J. Riedl. Polylens: a recommender system for groups of users. In *Proceedings of ECSCW '01*. Kluwer Academic Publishers, 2001.

[20] A. G. Parameswaran and H. Garcia-Molina. Recommendations with prerequisites. In *Proceedings of RecSys '09*. ACM, 2009.

[21] C.-K. Wong and M. C. Easton. An efficient method for weighted sampling without replacement. *SIAM Journal on Computing*, 9(1):111–113, 1980.

[22] M. Xie, L. V. Lakshmanan, and P. T. Wood. Breaking out of the box of recommendations: from items to packages. In *Proceedings of RecSys '10*. ACM, 2010.