Manipulation-Resistant Recommender Systems through Influence Limits

Paul Resnick
School of Information, University of Michigan
and
Rahul Sami
School of Information, University of Michigan

In this letter, we outline a new approach to modeling, analyzing, and combating manipulative attacks on recommender systems.

Categories and Subject Descriptors: I.2.6 [**Computing Methodologies**]: Artificial Intelligence—*Learning*

General Terms: Algorithms, Reliability

Additional Key Words and Phrases: Recommender systems, manipulation-resistance, shilling, information loss

Prediction markets reward or punish participants for the marginal changes they make to communal predictions. Online learning algorithms limit the influence of individual participants, and adapt these limits over time. In this letter, we summarize our recent work that combines these ideas in the realm of recommender systems, to make recommender systems robust against manipulation while making good use of information from genuine raters [Resnick and Sami 2007; 2008].

Recommender systems use collaborative filtering of past ratings to guide users to items they are likely to appreciate. The 'ratings' used in the collaborative filtering include explicitly provided feedback in the form of ratings or tags, as well as feedback that can be implicitly inferred by monitoring users' behavior such as browsing, linking, or buying patterns. Internet recommender systems have been deployed for a range of item categories, including books (e.g., Amazon.com), movies (e.g., Netflix), photographs (e.g. Flickr.com), websites (e.g., search engines, or social bookmarking sites such as del.icio.us). The recommendations provided by these systems can be of great value to certain entities, as evidenced by the multi-billion dollar search advertising market.

Users or organizations with a vested interest in having certain items recommended may attempt to manipulate the recommender system for their own profit. For ex-

Author's addresses: presnick@umich.edu, rsami@umich.edu. The work reported here was partially supported by the NSF under awards IIS-0812042, CCF-0728768, and IIS-0308006.

ample, the article recommender site Digg.com spawned a side market in which entities could pay to get their articles rated highly by a section of users [Newitz 2007]. This threat is exacerbated if, as on many online sites, an attacker can easily create multiple identities to carry out an attack. Attacks of this nature are called *sybil* or *shilling* attacks; many authors have noted that current collaborative filtering recommender systems are vulnerable to shilling attacks [Lam and Riedl 2004; O'Mahony et al. 2002], and proposed ways to combat this threat by distinguishing attack profiles from honest rater profiles [Chirita et al. 2005; O'Mahony et al. 2006; Mehta et al. 2007; Sandvig et al. 2007; Mehta and Nejdl 2008; Mobasher et al. 2007]. Another approach that has been proposed is to use monetary payments to create incentives for other raters to counteract the attacks [Bhattacharjee and Goel 2007]. In contrast, our approach does not require monetary transfers, and is provably manipulation-resistant (in the specific sense described below) without any assumptions about the number or percentage of honest raters.

Our objective is to bound the damage that an attacker can do with a fixed number of sybils; a secondary objective is to make effective use of information from honest raters. The first objective is primary because when damage is limited, attackers will have no incentive to attack. By contrast, online learning research typically combines both objectives by analyzing the total regret, which comes both from using false information and ignoring good information [Cesa-Bianchi and Lugosi 2006].

The central feature of our approach is to utilize the dynamic sequence of ratings instead of just a snapshot of rating profiles. This is analogous to a prediction market, in which a trader's profit depends on the market state (price) at the time she trades. Consider the recommendations made to a single target user T. (T can also be a group of users who receive identical recommendations.) The recommender system can track a natural measure of every rater j's contribution to the recommendations for T: For each item that j rated, we can track the change that j's rating caused to the prediction to target T. If T later rates the item, we can measure the impact that j's rating had on reducing or increasing the error in the recommender's prediction. We use a myopic measure of impact: we only assess the immediate change in the prediction to T after j's rating, and not indirect effects of j's ratings on future predictions after more ratings arrive. The greater *j's influence* on an item – the more that j's rating changed the prediction – the greater the measured impact j can have, positive or negative. The Influence Limiter algorithm [Resnick and Sami 2007] attenuates the influence on the next item such that the maximum negative impact from a rater is no more than its cumulative positive impact on all the previous items, plus a small constant c/n, where the parameters c and n are determined by the robustness requirement, as explained below. Thus, the total negative impact for any rater is never more than c/n.

A related area of research is the study of voting systems that are robust to shilling attacks [Conitzer 2008]. A recent proposal for preventing attacks, when there is a nonzero cost to create a shill, imposes a limit on each entity's change to the distribution of voting outcomes, analogous to our influence limits [Wagman and Conitzer 2008]. The key difference is that we seek to limit net damage rather than prevent damage on any single item; this allows us to give informative raters

more influence over time rather than setting a static limit, and thus, to ignore less information from genuine raters.

The Influence Limiter has two attractive strategic properties. First, it ensures that, regardless of the sequence of ratings, the net damage that the attacker can cause – the extent to which she can increase the error of the recommender by raising the predictions of poor items or lowering the predictions of good items – is bounded above by c. Formally, we define a recommender system to be (n, c)-robust if any attacker using no more than n fake identities cannot cause immediate damage, summed over all items, of more than c. For any given n and c, the Influence Limiter can be parameterized to be (n, c)-robust under mild technical assumptions about the attack strategies. Second, a rater seeking to maximize her ability to influence predictions would rate honestly, to make the recommender's predictions as accurate as possible. Thus, the introduction of influence limits does not create perverse incentives to distort the recommendations.

Resistance to manipulation is only one aspect of a recommender system's performance. Indeed, we could build a perfectly secure recommender simply by discarding all ratings, and making arbitrary recommendations. Thus, it is important to address the accuracy or informativeness of the recommender as well as its robustness. By limiting the ability of new, genuine, raters to influence predictions, the Influence Limiter does discard potentially useful information. We use an information-theoretic model of rating informativeness to quantify the information lost due to influence limits. In [Resnick and Sami 2007], we proved an upper bound on this information loss: If the Influence Limiter is set up to be (n, c)-robust, the total volume of information from a rater that is discarded is $O(\log \frac{n}{c})$. In other words, the additional error due to placing influence limits on a non-attacking rater, totaled over all items that he rated, is at most $O(\log \frac{n}{c})$.

The Influence Limiter thus trades off robustness and information loss. It is natural to ask if it is possible to design recommender systems that are robust against an attacker that can create arbitrarily large number of raters. Unfortunately, the answer is no. In [Resnick and Sami 2008], we prove that the tradeoff between robustness and efficiency in the Influence Limiter is essential: Any (n, c)-robust algorithm must discard $\Omega(\log \frac{n}{c})$ information from each genuine rater in the system in the worst case. The proof uses a simple family of random-noise attacks; the bound is derived by showing that it is not possible to distinguish an informed rater from random noise with high probability until a large number of ratings have been observed. A recommender must limit the influence of an attacking random guesser in order to be robust, and hence, it must also frequently limit the influence of an informed genuine rater. One of the consequences of this result is that it is impossible to build a useful recommender that is robust against attacks with an unbounded number of sybils: any such recommender must discard all information from genuine raters.

There are several interesting directions for future work. The first is to extend the approach to a non-myopic measure of damage, *i.e.*, to account for the effects of an attack rating on future predictions after more ratings come in. The myopic measure we currently use corresponds to a prediction market, in that a trader is held responsible for the prediction they make based on earlier trades as well as

their private information. Even in a market setting, it is possible (but difficult) for traders to execute non-myopic attacks, in which they trade in order to mislead future traders, and then profit off the subsequent errors. However, in a recommender system, this threat is more acute: the predictions are typically computed by applying a well-known collaborative filtering algorithm to the set of ratings, and so an attacker can easily anticipate the effect of his rating on future predictions. Thus, it is important to develop algorithms that guard against such non-myopic attacks.

Although the upper and lower bounds on information loss are asymptotically comparable, there is a constant factor difference between the bounds; new algorithms or tighter lower bounds would be interesting. There are also interesting research directions in safely combining information from multiple targets to determine the influence limits, and adjusting for selection bias. At a higher level, we believe that the technique we use, which builds on ideas from the prediction market literature as well as machine learning, could be useful in other mechanism design problems.

REFERENCES

Bhattacharjee, R. and Goel, A. 2007. Algorithms and incentives for robust ranking. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA '07)*.

Cesa-Bianchi, N. and Lugosi, G. 2006. *Prediction, Learning, and Games*. Cambridge University Press.

Chirita, P.-A., Nejdl, W., and Zamfir, C. 2005. Preventing shilling attacks in online recommender systems. In *WIDM 05*. 67–74.

Conitzer, V. 2008. Anonymity-proof voting rules. In *Proceedings of the Fourth Workshop on Internet and Network Economics (WINE'08)*.

Lam, S. K. and Riedl, J. 2004. Shilling recommender systems for fun and profit. In *Proceedings of WWW '04*. 393–402.

Mehta, B., Hoffman, T., and Fankhauser, P. 2007. Lies and propaganda:detecting spam users in collaborative filtering. In *Proceedings of IUI'07*.

Mehta, B. and Nejdl, W. 2008. Attack resistant collaborative filtering. In *Proceedings of ACM SIGIR 2008 (to appear)*.

Mobasher, B., Burke, R., Bhaumik, R., and Williams, C. 2007. Towards trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Transactions on Internet Technology 7,* 2, 1–40.

Newitz, A. 2007. I bought votes on digg. Wired Magazine. Available at *http://www.wired.com/techbiz/people/news/2007/03/72832*.

O'Mahony, M., Hurley, N., and Silvestre, G. 2002. Promoting recommendations: An attack on collaborative filtering. In *Proceedings of the 13th International Conference on Database and Expert System Applications*. Springer-Verlag, 494–503.

O'Mahony, M. P., Hurley, N. J., and Silvestre, G. C. M. 2006. Detecting noise in recommender system databases. In *Proceedings of the 2006 International Conference on Intelligent User Interfaces*. 109–115.

Resnick, P. and Sami, R. 2007. The influence limiter: Provably manipulation-resistant recommender systems. In *Proceedings of the ACM Recommender Systems Conference (RecSys07)*.

Resnick, P. and Sami, R. 2008. The informational cost of manipulation resistance in recommender systems. In *Proceedings of the ACM Recommender Systems Conference (RecSys08)*.

Sandvig, J., Mobasher, B., and Burke, R. 2007. Robustness of collaborative recommendation based on association rule mining. In *Proceedings of the 2007 ACM Conference on Recommender Systems*.

Wagman, L. and Conitzer, V. 2008. In *Proceedings of AAAI'08*. 196–201.