# User-Specific Decision-Theoretic Accuracy Metrics for Collaborative Filtering

**Giuseppe Carenini**
Computer Science Dept. University of British Columbia
2366 Main Mall, Vancouver, B.C. Canada V6T 1Z4
carenini@cs.ubc.ca

## ABSTRACT

Accuracy is a fundamental dimension for the effectiveness of recommender systems. Several accuracy metrics have been investigated in the literature. However, we argue, these metrics are not sufficiently user-specific. In previous work, we proposed accuracy metrics that take into account a user-specific pointwise decision threshold. In this paper, we present even more user-specific accuracy metrics that rely on the user utility function on the rating scale as well as on a user-specific sigmoid functional decision threshold.

### Keywords

Collaborative Filtering, Evaluation Metrics, Decision Theory.

## INTRODUCTION

A critical step in testing a recommender system is the choice of a set of evaluation metrics appropriate for the specific recommender task. However, it is only very recently that a coherent framework to support such a choice has been presented in the literature. In [7] Herlocker et al. critically discuss evaluation metrics that have been used in the past to test collaborative filtering (CF) recommender systems and propose several topics for future work.

Traditionally the most investigated and applied metrics have been measures of how accurately the system can predict the rating of items (*Accuracy Metrics*). [7] discusses similarity and differences among several popular accuracy metrics and shows how certain accuracy metrics are more appropriate for certain user tasks. For instance the ROC metric is more appropriate when the user wants to find good items and there is a clear binary characterization of items (as relevant/non-relevant). Furthermore, in [7], accuracy metrics are compared in an empirical testing which suggests that when applied on several different variations of the same CF algorithm all accuracy metrics appear to group in only three distinct classes.

Although accuracy is a fundamental dimension for recommendation effectiveness, it is not the only one. Other metrics explored in previous work and discussed in [7] include: (a) *Coverage* - the portion of the domain items for which the system can generate predictions; (b) *Learning Rate* - How the prediction ability of the system increases as more data is provided; (c) *Novelty and Serendipity* - Whether the recommended items are not known by the user and whether the user would probably not have discovered those items; (d) *Confidence* - How effectively the system is able to generate and express its confidence in the predicted ratings; (e) *User Evaluation* - How real users actually react to the system's recommendations when they interact with a recommender system in lab or field studies.

After a detailed survey of all current metrics for testing CF Herlocker et al. suggest the following guidelines for *applying* existing metrics and *developing* new metrics:

Researchers in CF should:

- choose the accuracy metrics that best match their assumed user task(s).
- consider in their choice the finding that accuracy metrics group in three equivalence classes.
- further develop and apply the other metrics besides accuracy (described above).
- develop comprehensive quality measures that effectively integrate accuracy metrics with the other metrics.

Although we recognize these as extremely valuable suggestions, in this paper we propose another goal for research on CF evaluation metrics and present some preliminary steps to achieve it.

We argue that more work is necessary to devise more adequate accuracy metrics. Remarkably, the only information about the user required by all the accuracy metrics currently

Workshop: Beyond Personalization 2005
*IUI'05*, January 9, 2005, San Diego, California, USA
http://www.cs.umn.edu/Research/GroupLens/beyond2005

used to evaluate CF systems are the user's true ratings. We believe that more informative metrics should take into account more user specific information, especially if this information can be easily acquired.

In previous work [1] we have proposed a first step in this direction, namely that adequate accuracy metrics should take into account a user specific pointwise threshold on whether to accept or refuse a recommendation. In this paper we move a step further, we argue that accuracy metrics should be even more user specific. In particular, first, in determining how much the user is gaining from following a recommendation, they should take into account the user utility function in the rating scale. Secondly, the user specific threshold should not be required to be a point. We claim that a sigmoid function would be a more adequate and more user-specific representation for the threshold.

In the remainder of the paper, we first report on our previous work on decision-theoretic user-specific accuracy metrics based on a pointwise user specific threshold. Next, in light of results from a user study and practical considerations, we reconsider two key assumptions on which our metrics were based. After that, we present our novel metric that does not rely on the two assumptions. We then discuss possible problematic aspects of our proposal and describe future work to address them.

## OUR PREVIOUS WORK ON DECISION-THEORETIC ACCURACY METRICS FOR CF

The Mean Absolute Error (MAE) is the most commonly used measure to evaluate the accuracy of CF algorithms. Let's assume the set $P_a$ with cardinality $n_a$ contains the ratings that the CF algorithm attempts to predict for current user $a$, then the MAE for that user is given as follows:

$$MAE_a = \frac{1}{n_a} \sum_{j \in P_a} |o_{a,j} - p_{a,j}|$$

where $o_{a,j}$ is user $a$'s observed rating for item $j$ and $p_{a,j}$ is user $a$'s predicted rating for item $j$. The MAE reported in CF evaluations is the average MAE for all users in the test set. Notice that the lower the MAE, the more accurate the CF algorithm is.

In [1] we criticize MAE because it relies on viewing the recommendation process as a machine learning problem, in which recommendation quality is equated with the accuracy of the algorithm's prediction of how a user will rate a particular item. This perspective is missing a key aspect of the recommendation process. The user of a recommender system is engaged in deciding whether or not to experience an item (e.g., whether or not to watch a movie). So, the value of a recommendation critically depends on how the recommendation will impact the user decision making process and only indirectly on the accuracy of the recommendation. For illustration, consider a user who will watch only movies whose predicted rating $p_{a,j}$ is greater than 3.5. Now, consider the following two predictions for that user:

*(i)* $p_{a,j} = 0$; when $o_{a,j} = 2$; (Absolute Error = 2)

*(ii)* $p_{a,j} = 3$; when $o_{a,j} = 4$; (Absolute Error = 1)

The Absolute Error in *(i)* is greater than the one in *(ii)*. However, in terms of user decision quality, *(i)* leads to a good decision because it entails that the user will avoid watching a movie that she would not have liked, while *(ii)* leads to a poor decision, because it entails that the user will miss a movie that she would have enjoyed.

The key point of this example is that in measuring the quality of a recommendation we should take into account the criterion used by the user in deciding whether or not to accept the recommendation. When the recommendation is provided as a predicted rating, a simple plausible criterion is a user specific threshold in the range of possible ratings. The user will accept a recommendation when the prediction is greater than the threshold.

More formally, let $\theta_a$ be a user specific threshold (in the range of possible ratings) such that:
$p_{a,j} \geq \theta_a \Rightarrow select(a, j)$

where, as before, $p_{a,j}$ is user $a$'s predicted rating for item $j$, and $select(a, j)$ means that user $a$ will select item $j$ (e.g., the user will watch the movie).

Then, the quality of a recommendation $p_{a,j}$, which we call User Gain (UG), can be defined as:

$$UG(p_{a,j}) = \begin{cases} o_{a,j} - \theta_a & \text{if } p_{a,j} \geq \theta_a \\ \\ \theta_a - o_{a,j} & \text{otherwise} \end{cases}$$

where, as previously defined, $o_{a,j}$ is user $a$'s observed rating for item $j$.

The first condition covers the situation in which, since the prediction is greater than the threshold, the user will decide to experience the item and will enjoy it to the extent that its true rating is greater than the threshold. The second condition covers the situation in which the user will decide not to experience the item. In this case, the user will gain to the extent that the item's true rating is smaller than the threshold.

Similarly to MAE$_a$, we can also define the active user Mean User Gain (MUG$_a$) as:

$$MUG_a = \frac{1}{n_a} \sum_{j \in P_a} UG(p_{a,j})$$

and MUG as the average MUG$_a$ for the test set

In [1] we also discuss how the revision of MAE leads to a revision of the ranked scoring (RS) metric, a less commonly used measure of recommendation quality which can be applied when the recommender presents a recommendation to the user as a list of items ranked by their predicted ratings. In [1], the revised user-specific version of RS is called $RS_a^{UG}$.

In this paper we take a second look at UG and notice that it also makes two rather unrealistic assumptions:

- In UG the gain/loss of a decision is measured as a difference between ratings (see UG def.) rather than as a difference between the user-specific utility (in decision-theoretic sense [2]) of those ratings. Doing so implies that the utility of different ratings is a linear function for all users (and consequently it is the same for all users).

- In UG the user decision threshold $\theta_a$ is a point on the rating scale. This excludes the possibility that for certain ratings a user may *sometimes* follow a recommendation with that rating and sometimes not.

To test the soundness of the first assumption we ran a user study. As for the second assumption we will criticize it on the basis of discussions with users.

### USER STUDY TO TEST UTILITY ASSUMPTION

In our user study participants filled out a questionnaire eliciting their utility functions for movie ratings (in the decision-theoretic sense). The questionnaire was completed by 15 participants (students and faculty at UBC).

This was based on the classic probability-equivalent (PE) procedure (see [2]), in which the utility of a rating $v$ is equal to the probability $p$ that makes the participant indifferent between:

- the gamble: watch a 5 rated movie with probability $p$; watch a 0 rated movie with probability $(1 - p)$.[1]

- and the certain outcome of watching a $v$ rated movie.

The outcomes of the utility elicitation process are summarized in Figure 1. The figure shows a box-and-whisker plot of the utility functions for the 15 participants in the study. It is clear that the utility of ratings varies considerably among participants [2].

In conclusion, this user study indicates that at least in the movie domain the assumption that all users have the same linear utility function on the rating scale is incorrect.

### POINTWISE vs. FUNCTIONAL DECISION THRESHOLD

We also argue that the second assumption of a pointwise threshold for the user decision is rather unrealistic. Although we do not have any evidence from formal experiments, in several discussions with users of movie recommenders it became clear that people follow deterministic strategies only for extreme ratings, but are more flexible for ratings in the middle of the scale. For instance, they would definitely (not) go to a movie rated in the interval (1-2) 4-5, but they may or may not go to a movie rated in the interval 2-4 [3]. To represent this kind of decision strategy we propose to represent the user decision threshold as a sigmoid function: $1/1+e^{-ax}$. A smooth and continuous thresholding function frequently used in AI, Economics and other fields. An example is shown in Figure 2 for the $0, \ldots, 5$ rating scale. This function may be interpreted as the probability of accepting a recommendation given a certain rating (these conditional probabilities do not have to sum up to 1).

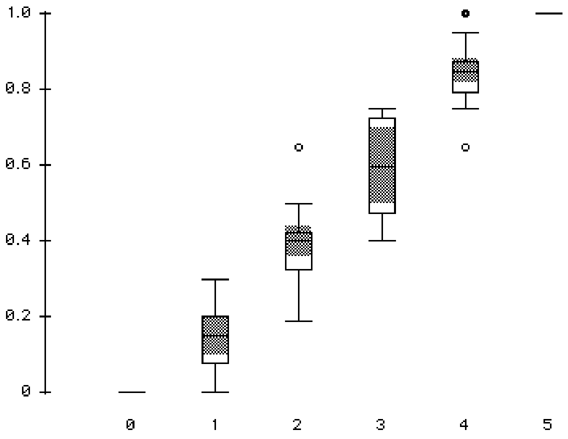


Figure 1: Box-whisker plot of the utility functions for the 15 participants in the study.

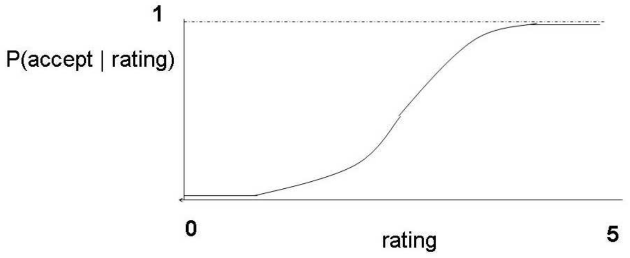


Figure 2: Sample sigmoid function for movie ratings: $1/1 + e^{-2(x-2.5)}$.

### NEW ACCURACY METRICS

In the previous two sections we have shown that two key assumptions underlying the specification of UG are rather unrealistic. In this section, we will define new accuracy metrics resulting from abandoning these assumptions.

[1] 5 and 0 are respectively the best and the worst possible ratings in the movie domain.

[2] The line in the middle of the box marks the median (the second quartile), the lower and upper extremes of the box mark the first and third quartile respectively. The two whiskers from each end of the box mark to the smallest and largest values (except for outliers which are marked by single points)

[3] We recently realized that empirical evidence for this kind of decision strategy is presented in [3]. When people have to map their movie ratings from a 1-5 scale to a binary scale (i.e., accept vs. do not accept the recommendation), their mappings strongly point to a sigmoid decision threshold. See Figure 2 in [3].

Let's begin with the assumption about the utility function. Since our user study indicated that users' utility function on the rating scale vary widely, we propose to redefine UG (and consequently RS$^{UG}$) into a new metric that explicitly considers the user specific utility functions.

Formally, if $u_a$ is the utility function for the active user on the rating scale and $\theta_a^u = u_a(\theta_a)$, a more refined UG could be defined as follows:

$$UG^u(p_{a,j}) = \begin{cases} u_a(o_{a,j}) - \theta_a^u & \text{if } u_a(p_{a,j}) \geq \theta_a^u \\ \\ \theta_a^u - u_a(o_{a,j}) & \text{otherwise} \end{cases}$$

And correspondingly more refined MUG$^u$ and $RS_a^{UG^u}$ could be defined, in which UG is substituted with UG$^u$.

As for abandoning the second assumption, we propose to revise $UG^u$ so that the user decision criterion for accepting a recommendation is not based on a user-specific pointwise threshold, but on a user-specific sigmoid function $sig_a$ which expresses the probability that a user will decide to experience an item with a given rating. Remember that $UG$ considered two possibilities (i) when the prediction is greater then the threshold, the user will decide to experience the item, otherwise (ii) the user will decide not to experience the item. With a sigmoid threshold, we have that a user given a predicted rating $p_{a,j}$ will decide to experience the item with probability $sig(p_{a,j})$ and decide not to experience the item with probability $(1 - sig(p_{a,j}))$. Therefore, we have to take into account both possibilities simultaneously in what formally is an Expected User Gain (EUG):

$$\begin{aligned} &EUG^{u-sig}(p_{a,j}) = \\ &\qquad [sig(p_{a,j}) * (u_a(o_{a,j}) - \theta_a^u)] + \\ &\qquad [(1 - sig(p_{a,j})) * (\theta_a^u - u_a(o_{a,j}))] \end{aligned}$$

And correspondingly more refined MEUG$^{u-sig}$ and $RS_a^{EUG^{u-sig}}$ could be defined, in which UG$^u$ is substituted with EUG$^{u-sig}$.

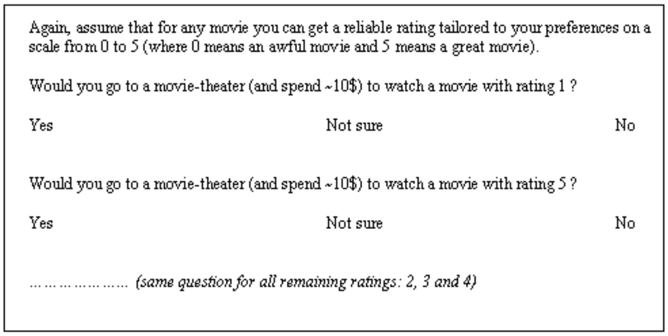
Notice that, in $EUG^{u-sig}$, the computation of the gain (the second factor in the products) still relies on a pointwise $\theta_a^u$. This is reasonable because it conceptually corresponds to the utility for the user of the resources invested in experiencing an item (e.g., time and money in the movie domain).

## OPEN ISSUES

Since our proposal is quite preliminary, there are several open issues that may lead to interesting discussions at the workshop.

### Elicitation of utility functions

- Will users be willing to go through the utility elicitation process? Presumably, it will depend on the number of ratings considered in the given domain.



Again, assume that for any movie you can get a reliable rating tailored to your preferences on a scale from 0 to 5 (where 0 means an awful movie and 5 means a great movie).

Would you go to a movie-theater (and spend ~10$) to watch a movie with rating 1 ?

Yes Not sure No

Would you go to a movie-theater (and spend ~10$) to watch a movie with rating 5 ?

Yes Not sure No

…………………… *(same question for all remaining ratings: 2, 3 and 4)*

Figure 3: Portion of the questionnaire to assess $\theta_a$ and $sig$

- Is there any alternative elicitation procedure? If some stereotypical utility functions can be acquired (in a given domain) then the problem of utility elicitation would be simplified to mapping a given user into the appropriate stereotype [6].

- Can we safely use utility functions as components of evaluation metrics?
  It is well known in decision theory that differences between values of utility functions (that express risk-attitudes) are meaningless unless the utility function is also a measurable value function [4]. This may or may not be the case ([8] pag. 132). So before applying $UG^{u-sig}$ in a domain, practitioners should make sure that the elicited utility functions are also measurable value functions.

### How to elicit the functional decision threshold?

- The sigmoid function $sig$ could be elicited by having the user fill out a questionnaire about her decision strategies in a given domain. Figure 3 shows the part of the questionnaire that we have designed for the movie domain. We have left the definition of the ratings as general as possible *"..where 0 means an awful movie and 5 means a great movie"* to preserve as much as possible the generality of the assessment. In interpreting the participant answers, the following schema could be applied. Let's call $max^{no}$ the highest rating for which the participant answered "no" and $min^{yes}$ the lowest rating for which the participant answered "yes". Then $\theta_a$ is assigned the midpoint of [$max^{no}$, $min^{yes}$] and $sig$ can be specified so that its inflection point is in $\theta_a$ and $sig(max^{no}) = sig(min^{yes}) \approx 1$

- Alternatively, the functional decision threshold could be acquired from the user behavior history. If the user was providing the system with feedback on whether she has decided to experience recommended items the system could simply construct the functional decision threshold by computing for each rating the frequency by which the user accepted recommendations with that rating.

## CONCLUSIONS AND FUTURE WORK

As CF recommender systems become more and more popular, there is a pressing need to develop effective evaluation metrics. Traditionally, accuracy metrics have been the most intensely investigated. In this paper, we reconsider popular accuracy metrics.

We noted that all accuracy metrics currently used do not sufficiently take into account the possibly highly user-specific decision process underlying the user interaction with a CF recommender system. To address this limitation we propose novel accuracy metrics based on the Expected User Gain measure. EUG is highly user-specific. It not only takes into account the user utility function on the rating scale in computing the user decision gain, but it also models the user decision criterion for accepting a recommendation as a sigmoid function expressing the probability that the user accepts a recommendation given a certain rating.

Once the open issues discussed in the previous section will be sufficiently clarified, we plan to apply our new metrics to compare existing CF algorithms. To do this, first we need to collect a new dataset in a domain that includes not only the user/rating matrix but also: (i) user specific $\theta_a$s; (ii) corresponding user specific $sig$ functions (using a questionnaire); (iii) user specific measurable utility functions. On this dataset, it will be possible to evaluate CF algorithms with $EUG^{u-sig}$. We are considering movies as our first domain because it is by far the most investigated domain for CF.

However, we would like to test the ideas presented in this paper in at least another domain besides movies. We plan to consider the joke recommendation domain [5]. For this domain, we will go through the same steps aiming to identify similarities and differences with respect to the movie domain. In particular, we are interested in verifying: (i) whether the utility function also varies widely across users in the joke domain (ii) whether in this domain it is more effective to acquire the functional decision threshold by means of a questionnaire or from the user behavior history.

As a long-term goal we plan to perform an evaluation of our approach in live systems. We hope that at the workshop we will be able to establish some form of collaboration with researchers who are running such systems.

## ACKNOWLEDGEMENTS

This research is being supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

## REFERENCES

1. G. Carenini and R. Sharma. Exploring more realistic evaluation measures for collaborative filtering. In *Proceedings of the 19th American National Conference of Artificial intelligence*, 2004.
2. R. T. Clemen. *Making Hard Decisions*. Belmont CA: Duxbury Press, 2nd edition edition, 1996.
3. D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl. Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of the Conf. on Human Factors in Comp. Systems CHI03*, p. 585 – 592, 2003.
4. S. Dyer and R. K. Sarin. Measurable multiattribute value functions. *Operations Research*, 27:810–822, 1979.
5. K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.
6. V. Ha and P. Haddawy. Toward casebased preference elicitation: Similarity measures on preference structures. In *Proc. of UAI*, pages 193–201, 1998.
7. J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1), 2004.
8. R. L. Keeney. *Value-Focused Thinking: A Path to Creative Decisionmaking*. Harvard University, 1996.