Evaluating Rank Accuracy based on Incomplete Pairwise Preferences

Brian Ackerman
Arizona State University
backerman@asu.edu

Yi Chen
Arizona State University
yi@asu.edu

ABSTRACT

Existing methods to measure the rank accuracy of a recommender system assume the ground truth is either a set of user ratings or a total ordered list of items given by the user with possible ties. However, in many applications we are only able to obtain implicit user feedback, which does not provide such comprehensive information, but only gives a set of pairwise preferences among items. Generally such pairwise preferences are not complete, and thus may not deduce a total order of items. In this paper, we propose a novel method to evaluate rank accuracy, *expected discounted rank correlation*, which addresses the unique challenges of handling incomplete pairwise preferences in ground truth and also puts an emphasis on properly ranking items that users most prefer.

Categories and Subject Descriptors

H.3.4 [**Storage and Retrieval**]: Systems and Software - Performance evaluation—*efficiency and effectiveness*

General Terms

Measurement, Performance

Keywords

Recommender systems, rank accuracy, pairwise preference, evaluation

1. INTRODUCTION

Recommender systems have proven their value in many applications where it is advantageous to personalize a user's experience. Currently, recommender systems power some of the world's most visited websites. Users visit websites like Amazon, Netflix, and Urbanspoon when they are looking for suggestions on items that may be of interest. These websites can suggest relevant items from a large collection of items and output a ranked list ordered by the predicted relevance to the user to aid in a user's decision making process.

Given the prevalence of various recommender systems, a critical question becomes their evaluation. Evaluation measures are used to compare different techniques of recommendation and give researchers an objective for optimization. Existing evaluation measures may take a set of user ratings as the ground truth [2, 4], or a total ordered list of items given by the user [9], with an option to support ties [3, 8]. Based on the weight given to items, there are two types of evaluation. Reference rank accuracy measures the similarity between the ranked list output by a system and the list of user's preferences, such as Spearman's ρ [8] and Kendall's τ [3]. There are also variations of area under the curve measures, such as the one used in [7], which are closely related to a general loss function. Utility rank accuracy measures give a stronger weight to the items that have a higher relevance according to the user, such as normalized discounted cumulative gain [2], which is widely adopted [1, 5, 6].

Ground truth often has to be obtained through explicit user feedback. In some applications, we may solicit users to give ratings on items and use such ratings as relevance scores, such as data obtained from Netflix and MovieLens. The rating data provides accurate user opinions on items, and can also provide a total order of user's preferences on items. However, such data requires explicit feedback, and may be considered as burdensome for some users.

On the other hand, there is great potential for a recommender system, and search engines, to leverage implicit user feedback which can also be used for evaluation. After a user issues a query, a set of items will be recommended. Among them, a user may select some to check for more details, and finally may choose one or more of the products to buy or places to visit. Such user interaction provides implicit user feedback on an item's relevance. We can consider the set of items purchased or visited to be preferred to the set of selected items, which in turn is preferred to the set of items that do not receive any user interaction. Alternatively, we may also measure the actual time that a user spends examining an item to infer finer grained preferences.

After collecting data in the form of implicit user feedback, we must also be able to evaluate recommendation techniques based on the data. Using the data, we may not have the actual rating of items, but a set of pairwise preferences among items. Note that generally *pairwise preferences may not be complete*. That is to say there may exist two items which a user's preference is unknown. For instance, we don't know a user's preference between an item that is presented to the user in response to a query and an item that is not shown to the user. In this case, a total order among a set of items cannot be deduced. In this paper, we only consider evaluating preferences that are not contradictory. We simply allow for the aggregation of preferences for different user sessions assuming the preferences are consistent. It is worth noting that considering contextual factors involved with a user session would also be advantageous, but is outside the scope

Table 1: Example Pairwise Preference Data

	Preferences
Ground Truth	$A \succ C$, $A \succ D$, $A \succ E$, $C \succ D$, $B \succ D$
Prediction	$C \succ A$, $C \succ B$, $C \succ D$ $A \succ E$, $B \succ E$, $D \succ E$

of this paper. These contextual factors could include the set of items the user's were recommended, items previously purchases or visited by the user, and how the items were displayed to the user. It is also possible to find cyclic preferences (e.g. $A \succ B$, $B \succ C$, $C \succ A$). However, we also do not consider these cases.

To evaluate recommender systems whose user preferences are obtained through implicit user feedback, we study evaluation methods for a general type of recommender system that takes a set of *possibly incomplete pairwise preferences* as input for training and for ground truth, and outputs a ranked list of items for recommendation. After examining why existing evaluation methods may fail, we proposed a novel rank accuracy measure, expected discounted rank correlation (EDRC), to handle incomplete pairwise preferences. EDRC handles two unique challenges in this setting, the lack of total ranking order on items, and possible unknown preferences between two items. EDRC also takes a discounted model that gives bigger penalty for wrong prediction on the more preferred items. To the best of our knowledge, this is the first attempt for designing an evaluation measure for ground truth and system prediction that is a set of incomplete pairwise preferences.

The remainder of this paper is outlined as follows: after we introduce the problem setting in Section 2, we briefly review two commonly used evaluation measures. Section 4 proposes a new method to measure rank accuracy for incomplete pairwise preferences, and Section 5 concludes the paper.

2. PROBLEM SETTING

In this work, we design an evaluation measure for a recommender system that takes ranking data in the form of pairwise preferences as input, trains on the ranking data to infer relevance values for each item in a test set, and then outputs a ranking order for the items based on their predicted relevance values. For the test set, we consider the ground truth to be a set of pairwise preferences collected from implicit user feedback.

We define a pairwise preference as a user's comparison between two items. We denote a pairwise preference as $A \succ B$ meaning item A is preferred to item B. Clearly, from a ranked list of items, we can derive a set of pairwise preferences. For example, we consider three items A, B, and C. The system may predict the order of these items to be A, B, C. Then the derived set of pairwise preferences is: $\{ A \succ B, A \succ C, B \succ C \}$. On the other hand, we may not always be able to derive a total order of items based on a set of pairwise preferences. For example, consider the set of pairwise preferences in the ground truth in Table 1 is insufficient to form a total order. For instance, we do not know the preference between A and B. We define a set of pairwise preferences to be *complete* if there exists a pairwise preference between every two items involved in the set, or a pairwise preference can be inferred. Otherwise, it is a set of *incomplete pairwise preferences*. As discussed in Section 1, incomplete pairwise preferences are common for user preferences that are obtained implicitly. This is true for both the training dataset and test dataset which contains the ground truth or user's preferences.

Table 2: Example Rating and Preference Data

$\mathcal{I}$	$rel(i)$	$index(i)$	$\widehat{index}(i)$
A	5	1	2
B	4	2	1
C	3	3	3
D	2	4	4

Since existing evaluation metrics are inapplicable for incomplete pairwise preferences in the ground truth, we propose a measure for rank accuracy where the ground truth is a set of possibly incomplete pairwise preferences. Since we can derive a set of pairwise preferences from the ranked list output, the core of the evaluation is to measure the similarity between two sets of pairwise preferences, one for system output, and the other from the user.

3. EXISTING EVALUATION METHODS

Before we discuss our newly proposed method of evaluation, we look at two existing methods, nDCG and AP correlation. For both methods we show a brief example and then discuss cases where each does not work for our problem setting.

3.1 Normalized Discounted Cumulative Gain

Cumulated gain-based evaluation was proposed by Jarvelin and Kekalainen in 2002 [2] to evaluate rank accuracy when the ground truth is a set of user ratings. The intuition is that it is more important to correctly predict highly relevant items than marginally relevant ones.

nDCG is formally defined in Equation 1. We denote $\mathcal{I}$ as a set of items suggested by the system, $rel(i)$ is the relevance of item i according to the user, $index(i)$ is the index of item i sorting the items based on the user's relevance score, and $\widehat{index}(i)$ is the index of item i sorted by the system's prediction.

$$nDCG = \frac{1}{Z} \cdot \left[\sum_{i \in \mathcal{I}} \frac{2^{rel(i)} - 1}{log_2(1 + index(i))} \right] \quad (1)$$

Z gives the maximum possible discounted cumulative gain if the items were correctly sorted, and is used as a normalization to ensure the result is between 0 and 1.

$$Z = \sum_{i \in \mathcal{I}} \frac{2^{rel(i)} - 1}{log_2(1 + \widehat{index}(i))} \quad (2)$$

For example, we look at the sample data in Table 2. If we look at item A, $rel(A) = 5$, $index(A) = 1$, and $\widehat{index}(A) = 2$. Below is the full sample calculation for $nDCG$.

$$nDCG = \frac{1}{Z} \cdot \left[\frac{2^4 - 1}{log_2(2)} + \frac{2^5 - 1}{log_2(3)} + \frac{2^3 - 1}{log_2(4)} + \frac{2^2 - 1}{log_2(5)} \right]$$

$$Z = \frac{2^5 - 1}{log_2(2)} + \frac{2^4 - 1}{log_2(3)} + \frac{2^3 - 1}{log_2(4)} + \frac{2^2 - 1}{log_2(5)}$$

This gives a value of .870 for the example.

nDCG assumes that we know an actual relevance for each item. It does not directly support the case when an item is preferred to another item, but the magnitude of the preference is unknown. When there is a total order on item preferences, we may assign rating scores with equal magnitude and then use nDCG. However, the absence of total order will result in such score assignments to be impractical, thus showing inapplicability of nDCG.

3.2 AP Correlation

AP (average precision) correlation (τ_{ap}) was proposed by Yilmaz et al. [9] as a modification to Kendall's τ which penalizes mistakes made for highly relevant items more than less relevant items. AP correlation finds the precision between two total orders at each index in the list and then takes the average of these values, as defined in Equation 3.

$$\tau_{ap} = \frac{2}{N-1} \cdot \left[\sum_{i \in \mathcal{I}} \frac{C(i)}{index(i) - 1} \right] - 1 \qquad (3)$$

N is the number of ranked items in the list and $C(i)$ is the number of items at an index less than $index(i)$ that are correctly ranked according to the ground truth. Consider the data in Table 2. For item A, $C(A) = 0$ because A comes after B in the prediction, but A is preferred to B in the ground truth. Below is a sample calculation.

$$\tau_{ap} \quad = \quad \frac{2}{4-1} \cdot \left[\frac{0}{1} + \frac{2}{2} + \frac{3}{3} \right] - 1 \quad = \quad \frac{1}{3}$$

AP correlation is measured on scale of -1 to +1, where -1 means the lists are in reverse order and +1 means the list are the same.

AP correlation assumes that each list, the ground truth and the system's prediction, gives a total order of items. There is no simple modification of AP correlation to support partial orders. Two challenges must be addressed when thinking about how to evaluate rank accuracy based on incomplete pairwise preferences. How does one assign a rank to an item when a total list cannot be constructed? In Equation 3, $index(i)$ is the rank index in the list, but when a total order is unknown, such as the sample data in Table 1, a new method is needed to give a rank index. Second, how does one consider the cases where a user's preference between items is unknown? In that example, we don't know user's preference between items A and B. How to evaluate a system that makes a predication $A \succ B$?

4. EXPECTED DISCOUNTED RANK CORRELATION

We propose *expected discounted rank correlation (EDRC)* to measure the similarity between two sets of pairwise preferences. Given a set of ground truth pairwise preferences from the user $\mathcal{G}$, and a set of predicted pairwise preferences output by the system $\hat{\mathcal{G}}$, EDRC calculates the expected correlation the two sets. Note that we may have user preferences on a large set of items based on many different user sessions. However, we only consider preferences relating items currently recommended by the system. That is, the set of items in $\mathcal{G}$ and $\hat{\mathcal{G}}$ are the same.

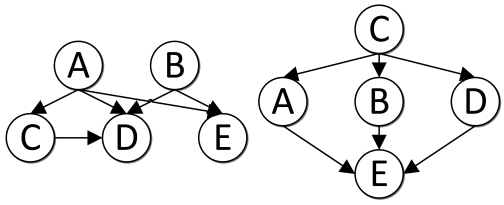

Figure 1: Topological Order based on Ground Truth (left) and System Prediction (right) in Table 1

Similar to AP correlation and nDCG, EDRC emphasizes preserving the order of the user's most preferred items and enforcing a smaller penalty for less preferred items. Different from nDCG whose ground truth is a set of relevance scores, the ground truth supported by EDRC is a set of pairwise preferences. Different from AP correlation which requires complete pairwise preferences, EDRC allows incomplete pairwise preferences.

As discussed in Section 3.2, the presence of incomplete pairwise preferences entails two challenges: the lack of total order among items in the ground truth, and unknown preferences between two items. Next, we look at the first problem.

Assigning Rank and Computing Weight. In the spirit of discounted gain, we want to give different *discount* (weight) for different items. If we know the rank of an item, $R(v)$, in the ground truth, we may set the weight linearly, logarithmically, or exponentially. However, the problem is how to compute the rank of an item given incomplete pairwise preferences. From a set of pairwise preferences as ground truth, we first derive a topological order among the items. Consider a graph where a vertex represents an item, and a directed edge represents a pairwise preference. The item corresponding to the source vertex is preferred to the item corresponding to the target vertex. In the following discussion, we use item and vertex interchangeably.

For example, the ground truth in Table 2 can be represented by the graph in Figure 1, where the fact that item A is preferred to item C is shown by a directed edge from vertices A to C. In this graph, an item is preferred to any item reachable following a directed path. For example, A is preferred to both C and D in the ground truth of Table 1.

Now we discuss how to leverage the graph to compute item rank. We initiate the rank of every vertex to be 1 and traverse the graph beginning for the start nodes. Whenever we follow an edge from a vertex u to v, we update v's rank to $max(R(v), R(u) + 1)$. Note that here we consider every user specified pairwise preference as equally important, and is assigned a unit weight of 1. Thus we keep the maximum score of node among the different paths that can reach this node from a start node. The procedure is defined in the procedure SetRanks($\mathcal{G}$) in Algorithm 1.

The ranks for the items in the graph of Figure 1 are as follows: $R(A) = 1$, $R(B) = 1$, $R(C) = 2$, $R(D) = 3$ and $R(E) = 2$. As we can see, nodes A and B have the same rank, since we don't know user's preferences between the two.

Algorithm 1 Setting Rank Value

SetRanks($\mathcal{G}$)
1: $\mathcal{S}$ = all nodes in $\mathcal{G}$ without an incoming edge
2: **for all** $v \in \mathcal{S}$ **do**
3: $R(v) \leftarrow 1$; Visit(v, 1)
4: **end for**

Visit(v, $rank_{in}$)
1: $R(v) \leftarrow \max(R(v), rank_{in} + 1)$
2: **for all** $w \in OUT(v, \mathcal{G})$ **do**
3: Visit(w, $R(v)$)
4: **end for**

Then we define the discount term, $D(v)$, which can take various forms depending on what type of discount is desired. For example, for a simple linear discount, $D(v) = R(v)$. For exponential discount, $D(v) = 2^{R(v)}$ and for logarithmic discount, $D(v) = log_2(1 + R(v))$.

Handling Unknown Preferences between Items. Another problem is computing the *score*, $C(v)$, of an item in the system's output in the presence of incomplete pairwise preferences in the ground truth. We propose the following formula to define the score of an item where $OUT(v, \mathcal{G}^+)$ is the set of outgoing edges from v in the transitive closure, $\mathcal{G}^+$, of $\mathcal{G}$, and W, the set of items that are more preferred or have an unknown relationship with v where $W = V(\mathcal{G}) \setminus [\{v\} \cup OUT(v, \mathcal{G}^+)]$. EP checks for the *expected score* regarding the relationship between v and all items in W.

$$C(v) = \sum_{w \in W} EP(v, w)$$

For a pair of items (v, w), suppose v preferred over w, that is, $v \succ w$ in $\mathcal{G}$. There are three cases. We have $v \succ w$ in $\hat{\mathcal{G}}$. In this case, $EP(v, w) = 1$. The second case, we have $w \succ v$ in $\hat{\mathcal{G}}$. Since the system prediction contradicts to the ground truth, we have $EP(v, w) = 0$. The third case, the system cannot predict a preference between v and w. Then we need to compute the expected score. By default we may assume there is 50% likelihood for $v \succ w$ and 50% likelihood for $w \succ v$. We then let $EP(v, w) = .5$. Alternatively, we may have a more accurate likelihood estimation based on collaborative filtering. Assuming we have a set of equally similar users, if 70% of the users have $v \succ w$ and 30% of similar users have $w \succ v$, then the expected score of $v \succ w$ is 0.7. For example, below are samples for C and EP based on Table 1.
$C(C) = EP(C, A) + EP(C, B) + EP(C, E) = 0 + .5 + .5 = 1$
$C(D) = EP(D, A) + EP(D, B) + EP(D, C) + EP(D, E) = .5 + .5 + 1 + .5 = 2.5$
$C(E) = EP(E, A) + EP(E, B) + EP(E, C) + EP(E, D) = 1 + 1 + .5 + .5 = 3$

Putting Things Together. Based on the discussion of how to compute a score for an item, $C(v)$, and the discount (weight) of an item, $D(v)$, we now put these together for an evaluation measure EDRC. We denote the set of all vertices in $\mathcal{G}$ without an incoming edge as $\mathcal{S}$.

$$EDRC(\mathcal{G}, \hat{\mathcal{G}}) = \frac{2}{Z} \cdot \left[\sum_{v \in V(\mathcal{G}) \setminus \mathcal{S}} \frac{C(v)}{D(v)} \right] - 1$$

Here, Z is a normalization factor to ensure the value is between +1 and -1.

$$Z = \sum_{v \in V(\mathcal{G}) \setminus \mathcal{S}} \frac{|W|}{R(v)}$$

Considering the example data in Table 1, we now show how to put together the sample calculation for EDRC using a linear discount method.

$$EDRC(\mathcal{G}, \hat{\mathcal{G}}) = \frac{2}{\frac{29}{6}} \cdot \left[\frac{1}{2} + \frac{2.5}{3} + \frac{3}{2} \right] - 1 = \frac{5}{29}$$

When both the ground truth and prediction is a complete set of pairwise preferences and the discount term is $D(v) = R(v) - 1$, the values for EDRC and AP correlation will be the same.

5. CONCLUSION

We consider a problem setting where the input from the user and the system's prediction are sets of possibly incomplete pairwise preferences. Based on this setting, we discuss the limitations of two evaluation methods, nDCG and AP correlation. Then we propose a new rank correlation metric, expected discounted rank correlation (EDRC), that compares two sets of pairwise preferences, one from the ground truth and one from the system prediction. This new measure has wide applications for evaluating recommender systems whose input data and ground truth are obtained from implicit user feedback.

Acknowledgements

This work was sponsored in part by an NSF CAREER Award IIS-0845647 and IIS-0915438. This work was also supported in part by an IBM faculty award.

6. REFERENCES

[1] L. Baltrunas, T. Makcinskas, and F. Ricci. Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the 4th ACM conference on Recommender systems*, 2010.
[2] K. Jarvelin and J. Kekalainen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, October 2002.
[3] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.
[4] R. Kumar and S. Vassilvitskii. Generalized distances between rankings. In *Proceedings of the 19th International World Wide Web Conference*, 2010.
[5] N. N. Liu and Q. Yang. Eigenrank: A ranking-oriented approach to collaborative filtering. In *Proceedings of the 31st Annual International ACM SIGIR Conference*, 2008.
[6] S.-T. Park and W. Chu. Pairwise preference regression for cold-start recommendation. In *Proceedings of the 3rd ACM conference on Recommender systems*, 2009.
[7] S. Rendle, C. Freudenthaler, Z. Gantner, and S.-T. Lars. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2009.
[8] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.
[9] E. Yilmaz, J. A. Aslam, and S. Roberston. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference*, 2008.