# Link Prediction Approach to Collaborative Filtering

Zan Huang, Xin Li, Hsinchun Chen
Artificial Intelligence Lab, Department of Management Information Systems, University of Arizona
McClelland Hall, 1130 East Helen Street
Tucson, Arizona, 85721-0108, USA
{zhuang, xinli, hchen}@eller.arizona.edu

## ABSTRACT
*Recommender systems* can provide valuable services in a digital library environment, as demonstrated by its commercial success in book, movie, and music industries. One of the most commonly-used and successful recommendation algorithms is *collaborative filtering*, which explores the correlations within user-item interactions to infer user interests and preferences. However, the recommendation quality of collaborative filtering approaches is greatly limited by the data sparsity problem. To alleviate this problem we have previously proposed graph-based algorithms to explore transitive user-item associations. In this paper, we extend the idea of analyzing user-item interactions as graphs and employ link prediction approaches proposed in the recent network modeling literature for making collaborative filtering recommendations. We have adapted a wide range of linkage measures for making recommendations. Our preliminary experimental results based on a book recommendation dataset show that some of these measures achieved significantly better performance than standard collaborative filtering algorithms.

## Categories and Subject Descriptors
H.1.2 [**Information Systems**]: User/machine systems– *Human information processing*; H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval – *Retrieval models*

## General Terms
Algorithms, Design, Experimentation

## Keywords
Recommender system, Collaborative filtering, Link prediction

## 1. INTRODUCTION
Recommender systems are widely used in many application settings to suggest products, services, and information items to potential users. At the heart of recommendation technologies are recommendation algorithms that take user/item attributes and user-item interactions (ratings, article browsing activities, book borrowing activities, etc.) as input to predict the unobserved level of match of a particular user-item pair. Collaborative filtering has been one of the most successful and well-studied recommendation algorithm, which only relies on the user-item interaction data to make recommendations. A standard user-based collaborative filtering algorithm first derives user neighborhoods by identifying similar users based on their overlapping interactions or similar ratings of common items. It then makes recommendations based on a user's neighbors' experiences. Despite its success, collaborative filtering is greatly limited by the data sparsity problem, where sparse user-item transactions do not support valid inference of user neighbors. For recommendations based on binary transaction data (examples in digital libraries include book borrowing and article browsing activities) we previously proposed to alleviate the sparsity problem by representing the transaction data as links in a bipartite graph containing user and item nodes [3]. Under this graph representation, the recommendation problem can be viewed as a task of selecting unobserved links for each user node, and thus can be modeled as a *link prediction* problem. In this study, we explore link prediction approaches developed recently in network modeling and prediction literature for transaction-based collaborative filtering recommendation.

We briefly discuss the network modeling literature and present six linkage measures adapted for recommendation in the next section. We then present a preliminary experimental study comparing link prediction approaches with standard collaborative filtering algorithms.


*JCDL '05*, June 7–11, 2005, Denver, Colorado, USA


## 2. LINK PREDICTION RECOMMENDATION ALGORITHMS
Network analysis is a new research methodology that has recently be applied to study a wide range of complex systems such as the Internet, WWW, social networks, and genetic interaction networks. Link prediction has been an important problem in network modeling and has recently been studied in social network, genetic interaction network, and literature citation network contexts [2, 5, 6]. In these studies certain linkage measures are defined to infer the potential for a future link to appear. In our study, we adapted six linkage measures employed in social network link prediction [5] for making recommendations based on a user-item interaction graph representation.

We represent the user-item interactions as links in a bipartite user-item graph $G$. Since we are studying the transaction-based collaborative filtering problem, the links are unweighted and the graph represents the complete information of the input data. Based on the topology of $G$ we calculate certain linkage measures $w(u, i)$ for each unconnected user-item pair $<u, i>$. These measures then serve as candidate scores for assessing the possibility for a link connecting $u$ and $i$ (a future transaction involving $u$ and $i$) for making recommendations.

Most existing linkage measures are proposed for unipartite graphs where links are allowed between any pair of nodes. We revised some measures to be meaningful within a bipartite graph. For a node $x$, we define $\Gamma(x)$ as the set of neighbors of $x$ in $N_h$. We also define $\hat{\Gamma}(x) = \bigcap_{c \in \Gamma(x)} \Gamma(c)$ as the set of neighbors of $x$'s neighbors.

The six linkage measures are listed below in two groups of neighbor-based (A) and path-based measures (B).

A1. *Common Neighbors*: In a unipartite graph, the number of common neighbors of $x$ and $y$ $|\Gamma(x)\bigcap\Gamma(y)|$ could represent the similarity of nodes. In our user-item graph, a user node is only connected with item nodes, thus $|\Gamma(x)\bigcap\Gamma(y)|$ will always be zero if $x$ and $y$ form a user-item pair. The bipartite version of the common neighbor measure in our study is $|\Gamma(x)\bigcap\hat{\Gamma}(y)|$.

A2. *Jaccard's Coefficient*: It measures the number of neighbors of both $x$ and $y$ compared to the number of nodes that are either $x$'s or $y$'s neighbors. In our context, this measure is adapted as $\left|\frac{\Gamma(x)\bigcap\hat{\Gamma}(y)}{\Gamma(x)\bigcup\hat{\Gamma}(y)}\right|$.

A3. *Adamic/Adar*: It computes features shared by objects, and defines the similarity between them as $\sum_{\text{z: feature shared by x,y}} \frac{1}{\log(frequency(z))}$ [1]. In our context, an object is a node and its features are its neighbors. The measure we use is $\sum_{z\in\Gamma(x)\bigcap\hat{\Gamma}(y)} \frac{1}{\log|\Gamma(z)|}$ after adaptation to accommodate for the bipartite characteristics.

A4. *Preferential Attachment*: $|\Gamma(x)|\cdot|\Gamma(y)|$ This measure relates the future interaction with the item's popularity and the users' activity level in the system.

B1. *Graph Distance*: This measure is defined as the length of the shortest path between pairs of nodes in $G$.

B2. *Katz$_\beta$*: Katz defined a measure that sums weights of all paths between two nodes exponentially damped by length [4]. In our context, this measure is defined as $\sum_{l=1}^{\infty}\beta^l\cdot|paths_{x,y}^{<l>}|$ where $paths_{x,y}^{<l>}$ is the set of all length-$l$ paths from $x$ to $y$.

The first three neighbor-based measures (A1-A3) involve intersection between neighbor sets, which are essentially similar to standard neighbor-based collaborative filtering algorithms. The Preferential Attachment measure has its roots in scale free network models and is related to the mechanism of network growth. The two path-based measures incorporate global network structure into individual linkage measures, which are expected to enhance recommendations under sparse data.

# 3. AN EXPERIMENTAL STUDY

We used a book sales dataset from a major Chinese online bookstore for our experimental study. This dataset covers 5 years of transactions of 2,000 customers, involving 9,695 books and 18,771 transactions. To evaluate recommendation performance, we adopted the top-N recommendation task. For each customer we recommended the top 10 books that he/she had not purchased previously ranked by individual linkage measures. For comparison purposes, we included standard user-based and item-based collaborative filtering algorithms as benchmarks [3]. We employed standard top-N recommendation quality measures including precision, recall, F measure, and rank score [3] to evaluate the accuracy, coverage, and ranking quality of the recommendations. The quality measures were derived by matching the recommendation lists against 20% of the actual purchase records we withheld for comparison purposes.

**Table 1. Experimental results: algorithm performance measures (boldfaced measures were not significantly different from the highest measure in at 5% level)**

| Algorithm | Precision | Recall | F Measure | Rank Score |
|---|---|---|---|---|
| Common Neighbors | 0.0181 | 0.0844 | 0.0279 | 3.0967 |
| Jaccard's Coefficient | 0.0060 | 0.0300 | 0.0096 | 1.2072 |
| Adamic/Adar | 0.0166 | 0.0805 | 0.0259 | 2.7252 |
| Preferential Attachment | **0.0258** | **0.1316** | **0.0404** | **8.6383** |
| Graph Distance | 0.0029 | 0.0143 | 0.0045 | 0.4726 |
| Katz(β=0.005) | **0.0261** | **0.1281** | **0.0407** | 5.8578 |
| User-based | 0.0122 | 0.0753 | 0.0202 | 4.9332 |
| Item-based | 0.0093 | 0.0443 | 0.0144 | 3.2146 |

The recommendation quality measures for each linkage measure and the benchmark algorithms are presented in Table 1. The Katz measure achieved the best performance, followed by Preferential Attachment, Common Neighbors, and Adamic/Adar measure. The results show that both path-based and neighbor-based approaches can significantly outperform the standard user-based and item-based algorithms. The poor performance of the Graph Distance measure may indicate the small world property of the user-item graph: any pair of nodes can be linked with a relatively short path. Overall, these results indicate that link prediction approaches and network analysis in general may be an important direction for improving existing collaborative filtering algorithms.

# 4. FUTURE DIRECTIONS

We plan to explore additional linkage measures from the network analysis and modeling literature. We also plan to develop a meta-framework for selecting appropriate network linkage measures to generate effective recommendations for specific datasets based on network topological and other structural properties.

# 5. ACKNOWLEDGMENTS

This work was supported in part by: NSF Information Technology Research, "Developing a collaborative information and knowledge management infrastructure", IIS-0114011, 2001-2004.

# 6. REFERENCES

[1] Adamic, L.A. and Adar, E. Friends and neighbors on the Web. *Social Networks*, *25,* 3 (2003), 211-230.

[2] Goldberg, D.S. and Roth, F.P. Assessing experimentally derived interactions in a small world. *In Proceedings of the National Academy of Sciences USA*, *100*, 8 (2003), 4372-4376.

[3] Huang, Z., Chen, H. and Zeng, D. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Transactions on Information Systems*, *22*, 1 (2004), 116-142.

[4] Katz, L. A new status index derived from sociometric analysis. *Psychometrika*, *18,* 1 (1953), 39-43.

[5] Liben-Nowell, D. and Kleinberg, J. The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, 2003, 556-559.

[6] Popescul, A. and Ungar, L.H. Statistical relational learning for link prediction. In *Proceedings of the Workshop on Learning Statistical Models from Relational Data*, 2003.