

Research Statement – Sean M. McNee

Over the past four years, I have explored how users perceive recommendation lists generated by different recommender algorithms, a serendipitous discovery from my research in generating recommendations for peer-reviewed research papers. In general, my research interests lie at the convergence of human-computer interaction, machine learning, information seeking/retrieval, and library and information sciences. Specifically, I am fascinated by personalization and recommender systems, the ability to generate and present personalized bits of information to users in a way they can digest, interpret, and use.

While at the University of Minnesota, I have been fortunate enough to collaborate with a variety of people, working in a variety of different roles with professors, researchers, and students from three continents. This has allowed me to follow four diverse lines of research:

1. Understanding new users and issues of recommendation confidence in MovieLens, an online collaborative filtering-based (CF-based) recommender system
2. Performing a full usability assessment of a new version of MovieLens, a recommender system I helped design and build.
3. Using recommender systems to generate recommendations for peer-reviewed research papers, including evaluations of both CF and content-based algorithms
4. Re-examining the recommendation process from an end user's perspective, through which I propose Human-Recommender Interaction theory (HRI) and a new set of recommender algorithm metrics

1. The New User Experience and Recommendation Confidence

A difficult problem for recommender systems is how best to learn about a new user. Users must provide information to the system before receiving personalized recommendations. The question is how to gather this information from users to ease user effort and maximize recommendation quality—how to optimize the *new user experience*. In [Rashid 2002], we studied six techniques that collaborative filtering recommender systems can use to learn about new users. These techniques select a sequence of items for the collaborative filtering system to present to each new user for rating. The techniques included:

- Entropy-based technique to select items of high information value
- Popularity technique using aggregate statistics to select items a user is likely to have an opinion about
- Balanced technique (formed by combining entropy and popularity techniques) that seek to maximize the expected number of bits learned per presented item
- Item-based collaborative filtering recommender to personalize items presented

We studied these techniques thru offline simulation experiments and thru a live experiment on MovieLens involving over 300 users. We found the following results: entropy-based techniques worked when users were able to rate the items presented, which rarely happened. Popularity techniques proved popular with users, but suffered from poor prediction accuracy. Personalized techniques did not perform well either, dragging users into a “similarity well” of highly similar items. This led to imbalanced user models and poor prediction accuracy. The balanced technique, on the other hand, proved to be an excellent compromise, providing users with a good new user experience.

Take Away Message

In a CF-based recommender system, the choice of learning technique used to present items for new users to rate significantly affects the new user experience, both in terms of user effort and prediction accuracy. Balanced approaches, which maximize the expected number of bits of information per item, dominated over other non-personalized approaches. Personalized approaches are haphazard, due to a lack of information about a user.

Building on this work, in [McNee 2003a] I focused on exploring the effects of letting the user *actively participate* in choosing the items used during the new user experience. Instead of only allowing the system present items for the new user to rate, we wanted users to both tell us the items they wished to rate. We compared three interfaces to elicit information from new users: have the system choose items to rate (using the best algorithm from the above work), asking users to choose items themselves to rate, and a mixed interface combining the two other methods. Figure 1 shows an image of the mixed interface. The two pure interfaces contained the appropriate half of the mixed interface. This experiment was carried out on MovieLens involving 163 new users.

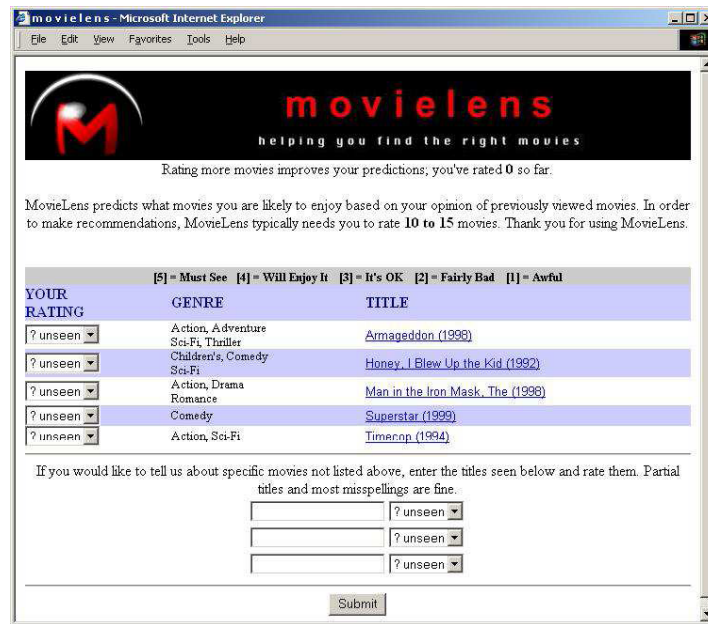


Figure 1: The Mixed Interface from [McNee 2003a].
In the top half, users were presented movies to rate. In the bottom half, users could enter movie titles and associated ratings.

We found the following results:

- Both pure interfaces had equal *perceived* user effort, even though users took much longer to when using a user-controlled interface.
- Both pure interfaces generated high quality user models, leading to high prediction accuracy
- The user-controlled interface had the largest dropout rate, but also had the most users return to the system and rate more items, both in relative and in absolute numbers. We believe these users had a higher loyalty to the system.
- The mixed interface was not a reasonable compromise, with the worst prediction accuracy and lowest loyalty rates of the three interfaces

Take Away Message

Forcing new users to be active participants in new user experience creates a higher initial burden than other possible interfaces. This causes more people to drop out of the signup process, but those who stay are more loyal to the system. The interface selection depends on the goals of the recommender.

Third, I was curious to know how introducing a confidence metric in to a recommender would affect different user populations (i.e. new vs. experienced users). In [McNee 2003b] we introduced a confidence metric into MovieLens and provided a different amount of training on this metric to both new and experienced groups of users, with 223 total users participating. An example of this metric is shown in Figure 2, with dice representing low confidence or a 'risky' recommendations. We found that training improved use of the confidence display compared to no training. New users were less likely to notice, understand, and use the confidence display than experienced users. On the other hand, providing training about a confidence display to experienced users greatly reduced user satisfaction in the recommender system because it altered their 'worldview' of how MovieLens worked.



Recent DVDs	
1. Beautiful Mind, A (2001)	★★★★★
2. Red Beard (Akahige) (1965)	★★★★★ 
3. From Hell (2001)	★★★★★
4. Traffic (2000)	★★★★★
5. Horse's Mouth, The (1958)	★★★★★ 

Figure 2: The confidence metric from [McNee 2003b]
More dice next to a recommendation implied more 'risk'.

Take Away Message

Users often request additional information about the recommendations they receive, such as prediction confidence. Reactions to introducing such metrics into an existing system vary based on user population. New users did not notice the interface, whereas the addition of a confidence metric alienated some existing users, breaking their worldview of how the recommender worked.

2. Architecting and Evaluating the Usability of MovieLens

For my M.S. project, I redesigned the user interface for and performed a full usability analysis on the MovieLens movie recommender system. This process included the use of rapid prototyping with an iterative design, heuristic analysis, cognitive walkthroughs, and a full usability study involving eight subjects in a fully equipped usability laboratory. Figure 3 provides a visual comparison between the two interfaces. This redesign faced significant challenges due to the introduction of new features, including an advanced search interface, movie buddies with whom you can get group recommendations, and the ability to 'save' a search to execute in the future. This had to occur without sacrificing the main goal of MovieLens: to rate movies and receive recommendations. Even when following such user-centered design principles, our usability study revealed several shortcomings including the use of buddies and the use of our 'tabbed' interface.

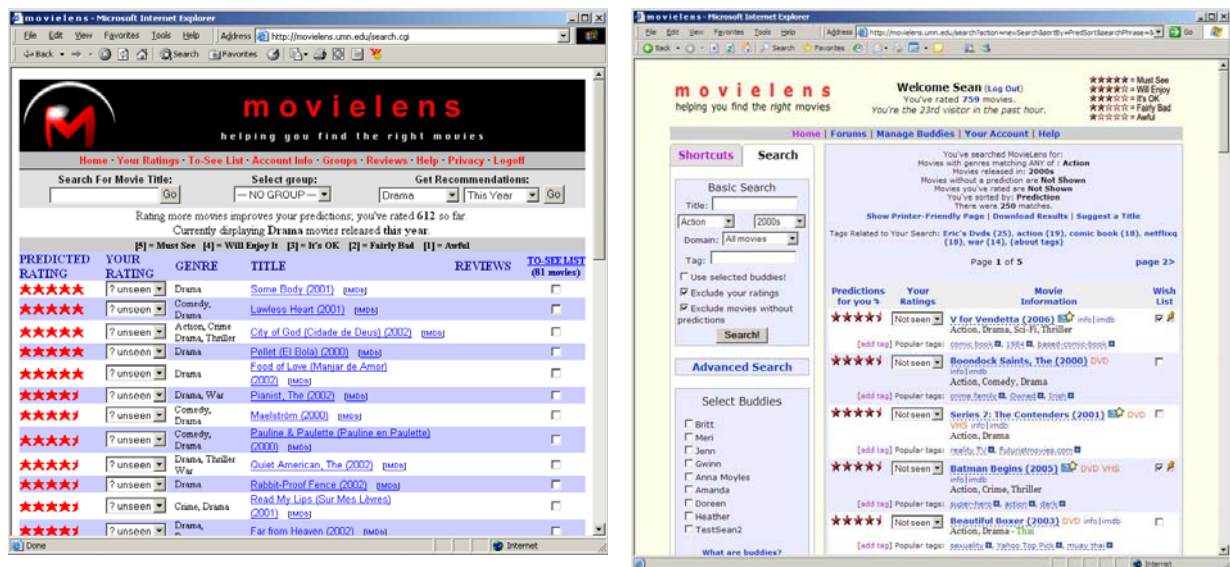


Figure 3: The two MovieLens interfaces. Old version on the left, new version on the right

Beyond designing and testing the interface, I also helped with the design and implementation the entire MovieLens platform. This version of MovieLens was released in May 2003, and is accessed by over 1,000 users each week.

Take Away Message

User-centered design processes can greatly enhance the usability of research-centered recommender systems. Yet, even the most detailed user-centered design is no substitute for real users.

3. Recommending Research Papers

In my first work in this area [McNee 2002], I explored the use of collaborative filtering to recommend research papers, to generate suitable additional references for a target research paper. As far as I am aware, this was the first research that built a system to study collaborative filtering in this domain. This system was based on a novel mapping of citation data to the CF ratings matrix. I investigated six algorithms for selecting citations, evaluating them through offline simulation experiments against a database of over 186,000 papers from ResearchIndex. I also performed an online experiment with over 120 users to gauge user opinion of the effectiveness of the algorithms and of the utility of such recommendations for common research tasks. I found large differences in the accuracy of the algorithms in the offline experiment, especially when balanced for coverage. Where as CF-based algorithms had high coverage, content-based algorithms had quite poor coverage—they demonstrated a 'cherry-picking' behavior, generating high quality recommendations for a small subset of input data. In the online experiment, users felt the

recommendations were useful and of high quality; users were happy when receiving one good recommendation in a list of five. More surprising, online experiment results also showed that users preferred different algorithms for different usage scenarios. For example, users stated that CF-based algorithms generated novel recommendations.

Take Away Message

By mining the citation network between research papers, collaborative filtering algorithms can generate high quality recommendation in this domain, without requiring user opinions or ratings of the items. Users felt recommendation in this domain were useful and of high quality. Moreover, users stated that specific recommender algorithms were better suited to different usage scenarios.

In [Torres 2004], we continued this line of research by introducing a set of hybrid recommender algorithms that combine collaborative filtering and content-based filtering for recommending research papers. See Figure 4 for an example of one hybrid algorithm we used. We evaluated our algorithms through offline experiments on a database of 102,000 research papers from ResearchIndex, and through an online experiment with 110 users. We developed separate English and Portuguese versions of the interface and specifically recruited American and Brazilian users to test for cross-cultural effects.

We found the following results:

- Not all hybrid recommenders generated high quality recommendations. “Chained” algorithms in which the results of one recommender were used as input to a second performed worse than fusion algorithms or pure algorithms
- We provided additional confirmation that different algorithms are better suited to different scenarios, with our users stating that CF algorithm generate novel recommendations where as content-based algorithms generate authoritative recommendations
- Students and novices had higher satisfaction from a recommender than professors and other research professionals
- Trend analysis suggested a cross-cultural effect, with Brazilian happier than American, but was not statistically significant

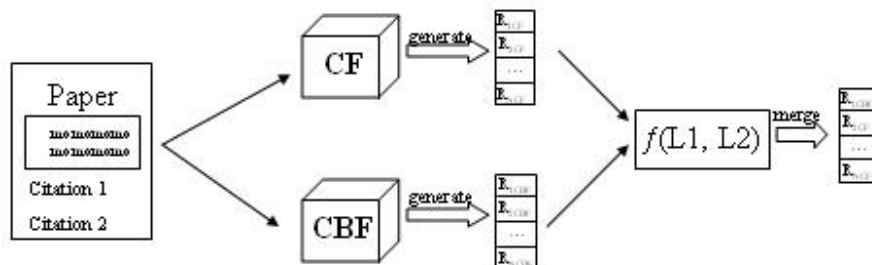


Figure 4: Our Fusion recommender algorithm. It merged lists using a voting algorithm

Take Away Message

Hybrid recommender algorithms combining CF- and content-based algorithms can generate high quality research paper recommendations. The usefulness of these recommendations varied across user populations with novices and students finding the system more useful than professors and research professionals. Finally, users felt that different algorithms generate qualitatively different recommendation lists.

Based on the results of these two papers, I started viewing the recommendation list as the important unit for judging the user experience in a recommender and began looking for list-based metrics to evaluate recommender algorithms. As a result, in [Ziegler 2005], we introduced an approach to diversify recommendation lists and a metric to measure list similarity. Specifically, we presented topic diversification, a novel method designed to balance and diversify personalized recommendation lists in order to reflect the user’s complete spectrum of interests. Further, we introduced the intra-list similarity metric to assess the topical diversity of recommendation lists. This work uses recommendation lists as entities rather than focusing on individual recommendations. We evaluated our method using book recommendation data from BookCrossing, including offline simulation experiments on over 360,000 ratings and an online user study involving more than 2,100 subjects. We found that diversification effect varied across recommender algorithms. Though being detrimental to average accuracy, we showed that our diversification method improves user satisfaction with recommendation lists, in particular for lists generated using item-based

collaborative filtering. Not a large change in diversity was required: the maximal effect occurred when 3 items out of 10 on a list were changed. In other words, small changes have a high impact.

Take Away Message

Users view recommendation lists as entities; results show list usefulness depends on more factors than only the usefulness of the items on the list. Depending on the user need, users were more satisfied with a diverse recommendation list, even at the expense of recommendation accuracy. Users noticed and appreciated small changes in recommendation lists.

4. Re-evaluating the Recommendation Process

Informed by this research, and after a careful review of the recommender systems, machine learning, user-centered design, and information seeking literature, I propose the following thesis:

In a recommender system, selecting and tuning the appropriate recommender algorithm for both the user and the user’s current information seeking task will generate a more useful recommendation list than a generic or un-tuned algorithm.

In support of this thesis, I present three research components: a theory component, an offline simulation component, and a user study component. In detail, these components are:

1. Human-Recommender Interaction theory (HRI), a novel framework for mapping recommender algorithms to information seeking tasks via metrics
2. A family of new algorithm metrics and a series of offline simulation experiments categorizing the differences between recommender algorithms
3. An online user evaluation with over 130 users to validate my approach

Figure 5 shows an overview of my proposed solution. In this process model, users and their information seeking tasks are linked to recommender algorithms through HRI and a set of recommender metrics. By describing a user task in the language of HRI, the most appropriate recommender algorithm can be selected based that algorithm’s performance on the relevant metrics. The model flows in both directions; tasks can be mapped to algorithms, and algorithms can be mapped to tasks.

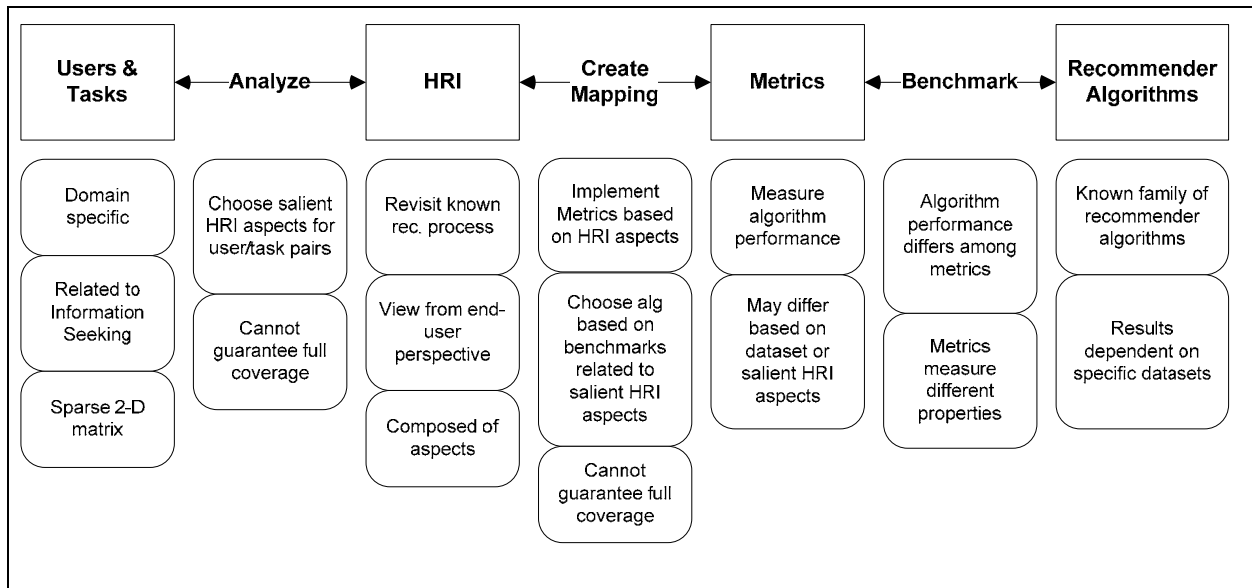


Figure 5: The HRI Process Model

Human Recommender Interaction Theory

To understand users in a recommender, I developed Human-Recommender Interaction theory (HRI) [McNee 2006b]. HRI is a way to understand the recommendation process from the end user’s perspective, providing an intermediary language to describe how end users view the recommendation process and their information seeking

needs, see Figure 6. There are three Pillars to HRI, representing the three parts of the recommendation process affecting users. First is the Recommendation Dialogue, the specific interaction where users give ratings and receive recommendations. Second is the Recommender Personality, in which the user prescribes personality characteristics to a recommender as part of repeated interactions over time. Third is the Information Seeking Task, the context in which a user comes to a recommender.

Within each Pillar are a set of Aspects. These terms form the language by which a user can express how the recommender should change to better meet his need. This novel expression of user opinion can be interpreted by recommender system designers and implementers (as well as the system itself!) to improve the recommendations a user receives. This process is grounded using a series of recommender algorithm metrics. These metrics demonstrate and categorize the qualitative differences in recommendation lists across recommender metrics, and through the language of HRI, these differences can be matched to user information seeking tasks.

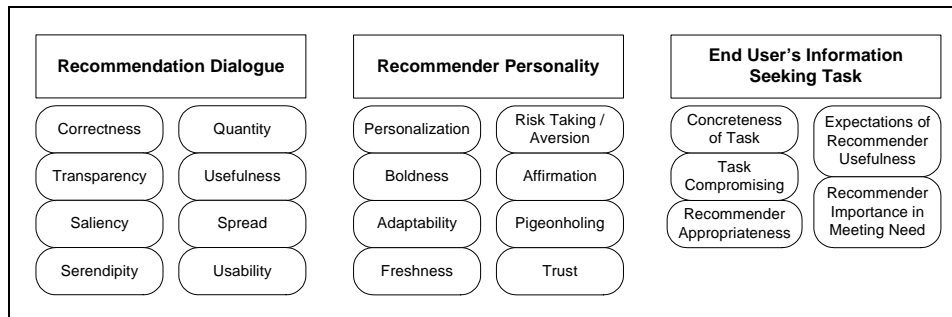


Figure 6: The Three Pillars of HRI. Each Pillar contains a set of Aspects

Take Away Message

Looking the recommendation process from the end user's perspective highlights the differences between recommender designer and end users. Human-Recommender Interaction theory is an intermediary language that both a recommender and its users can use to help generate more meaningful and useful recommendations. Using the HRI Process Model, user information seeking tasks can be mapped to recommender algorithms via the HRI descriptions and a family of metrics.

New Recommender Metrics and Simulation Experiments

Focusing on the recommendation list as the preferred unit of measure, I developed a series of new recommender metrics. These metrics are meant to work in tandem with existing predictive accuracy and decision support metrics, as the existing metrics are not rich enough alone to categorize algorithms [McNee 2006a]. In addition to the Intra-List Similarity metric described above, I created more new metrics, such as:

- ***Boldness***
The ratio of seen 'extreme' recommendations from a recommender algorithm to the expected number of 'extreme' recommendations
- ***Adaptability***
A measure of the ability of a recommender to change a recommendation list in response to changes in the user's rating profile
- ***Personalization***
A measure of the item overlap across all recommendation lists for different users, or compared against an external authority measure, such as item popularity

Running these metrics against several recommender algorithms including collaborative filtering and content-based algorithms, I found many interesting results. Of interest is the division between CF and statistical recommenders to the content-based recommenders. Content-based filtering demonstrated very different behaviors from the other algorithms across all metrics. This result underscores the limitations of current digital libraries using content-based search systems. Additionally, there were interesting results between user- and item-based collaborative filtering algorithms. As neighborhood size increased, the two algorithms diverged: user-based becomes more popular and less adaptable; Item-based CF is the opposite, becoming more adaptable as neighborhood sizes increase. In general, user-based is more volatile as neighborhood sizes change.

Take Away Message

Recommender algorithms differ from each other across many dimensions in terms of the recommendation lists they generate. Only by defining a set of metrics, each of which explores a different property, can we better understand the subtle differences in recommendation lists. The differences between collaborative filtering and content-based algorithms suggests that the information seeking interfaces to current digital libraries can be enhanced by adding a collaborative filtering-based search interface.

User Study

There are many potential pitfalls when tailoring recommendation algorithms to match user information seeking tasks, including not knowing what tasks to support, generating recommendations for the wrong task, or even failing to generate any meaningful recommendations whatsoever. To validate our findings from our HRI-based offline metric calculations, I perform a detailed user study with over 130 users to understand differences between recommender algorithms through an online survey of paper recommendations from the ACM Digital Library. I found that pitfalls are hard to avoid. Two of the algorithms, a naïve Bayes classifier and probabilistic latent semantic analysis algorithm generated ‘atypical’ recommendations—recommendations that were unrelated to their input basket. Users reacted accordingly, providing strong negative results for these algorithms. Our simulation testing did not reveal these problems with the generated results, as the recommendation lists were composed of both expected and atypical items. Results from our ‘typical’ algorithms show some qualitative differences. Specifically, we have further support that CF-based recommendations are considered ‘novel’, whereas content-based recommendations are considered ‘authoritative’. Since users were exposed to multiple algorithms in the study, these results may be affected by the atypical algorithm performance.

Take Away Message

Don’t look stupid. Recommenders that generate nonsensical results were not liked by users, even when the nonsensical recommendations were intermixed with meaningful results. Further, the evaluation must be done with real users, as current predictive accuracy metrics cannot detect these problems.

Future Research Plans

There are several immediate next steps in this research, including the design of an interface to support recommendation browsing, applying recommender algorithms to more complex user tasks, and expanding into other digital libraries. GroupLens Research recently received a grant from the NSF to work on this line of research.

There are several themes arising from this research that appeal to me. One such theme is how users approach the information seeking process, especially online. For example, this includes the use of personalization to increase trust, the integration of searching and browsing interfaces, and the application of recommender technologies to networks of artifacts. While my current research centers on personalization and collaborative filtering recommenders, I am interested in broader range of topics. In particular, I am quite interested at the point where human-computer interaction meets machine learning and information seeking theory—everything from information visualization to intelligent query processing to social networks. There are many exciting research questions in this space, and I would love to explore them from a variety of angles using all of the tools at my disposal, produce exceptional research, and have an impact with my results.

Refereed Full Papers

Sean M. McNee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, and John Riedl. “On the Recommendation of Citations for Research Papers.” *ACM CHI Letters* 4(3), CSCW 2002, ACM Conference on Computer-Supported Cooperative Work, New Orleans, November 2002, pp. 116-125. (Acceptance rate: 20%)

Sean M. McNee, Shyong K. Lam, Joseph A. Konstan, and John Riedl. “Interfaces for Eliciting New User Preferences in Recommender Systems.” In Proceedings of the 9th International Conference on User Modeling (UM 2003), Springer LNAI 2702, Johnston, PA, June 2003, pp. 178-188. (Acceptance rate: 25%)
This paper received the James Chen Best Student Paper award from UM 2003.

Sean M. McNee, Shyong K. Lam, Cathy Guetzlaff, Joseph A. Konstan, and John Riedl. “Confidence Metrics and Displays in Recommender Systems.” In *Proceedings of INTERACT '03 IFIP TC13 International Conference on Human-Computer Interaction*, Zurich, Switzerland, September 2003, pp. 176-183. (Acceptance rate: 34%)

Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K. Lam, Sean M. McNee, Joseph A. Konstan, and John Riedl. "Getting to Know You: Learning New User Preferences in Recommender Systems." In *Proceedings of the 2002 International Conference on Intelligent User Interfaces (IUI 2002)*, San Francisco, CA, January 2002, pp. 127-134. (Acceptance rate: 30%)

Roberto Torres, Sean M. McNee, Mara Abel, Joseph A. Konstan, and John Riedl. "Enhancing Digital Libraries with TechLens+." In *Proceedings of the Fourth ACM/IEEE Joint Conference on Digital Libraries (JCDL 2004)*, Tucson, AZ, June 2004, pp. 228-237. (Acceptance rate: 30%)

Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Larson. "Improving Recommendation Lists Through Topic Diversification." In *Proceedings of the Fourteenth International World Wide Web Conference (WWW 2005)*, Chiba, Japan, May 2005, pp. 22-32. (Acceptance rate: 14%)

Sean M. McNee, Nishikant Kapoor, Joseph A. Konstan. "Don't Look Stupid: Avoiding Pitfalls when Recommending Research Papers." In *Proceedings of the 2006 ACM Conference on Computer Supported Cooperative Work (CSCW 2006)*, Banff, Canada, November 2006. [To Appear] (Unknown acceptance rate)

Short papers, Demonstrations, and Other Publications

Joseph A. Konstan, Nishikant Kapoor, Sean M. McNee, and John T. Butler. "TechLens: Exploring the Use of Recommenders to Support Users of Digital Libraries". A Project Briefing at the *Coalition for Networked Information Fall 2005 Task Force Meeting*, Phoenix, AZ, December 2005.

Joseph A. Konstan, Sean M. McNee, Cai-Nicolas Ziegler, Roberto Torres, Nishikant Kapoor, and John Riedl. "Lessons on Applying Automated Recommender Systems to Information Seeking Tasks". A Nectar paper in the *Proceedings of AAAI-06: the Twenty-First National Conference on Artificial Intelligence (AAAI- 06)*, Boston, MA, July 2006. [To appear]

Sean M. McNee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, and John Riedl. "TechLens: A Recommender for Computer Science Research Papers." A Demonstration in Conference Supplement of *CSCW 2002, ACM Conference on Computer-Supported Cooperative Work (CSCW 2002)*, New Orleans, LA, November 2002, pp. 115-118.

Sean M. McNee, John Riedl, and Joseph A. Konstan. "Being Accurate is Not Enough: How Accuracy Metrics have hurt Recommender Systems". A Work-In-Progress paper in the *Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems (CHI 2006)*, Montreal, Canada, April 2006, pp. 1103-1108.

Sean M. McNee, John Riedl, and Joseph A. Konstan. "Making Recommendations Better: An Analytic Model for Human-Recommender Interaction". A Work-In-Progress paper in the *Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems (CHI 2006)*, Montreal, Canada, April 2006, pp. 1097-1101.

Papers in Progress

Sean M. McNee, John Riedl, and Joseph A. Konstan. "Breaking the Curse of Accuracy: Using Human-Recommender Interaction to Meet User information Needs in Recommender Systems." In preparation, 55 pages.

Sean M. McNee, Shilad Sen, Joseph A. Konstan, and John Riedl. "Examining Recommender Algorithm Performance, A Multi-Dataset Approach". In preparation, 10 pages.

Sean M. McNee, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, and John Riedl. "Understanding and Improving the New User Experience in Recommender Systems". In preparation, 30 pages.