# Correlating Perception-Oriented Aspects in User-Centric Recommender System Evaluation

Alan Said, Brijnesh J. Jain, Andreas Lommatzsch, Sahin Albayrak
Technische Universität Berlin
Berlin, Germany
{alan, jain, andreas, sahin}@dai-lab.de

## ABSTRACT

Research on recommender systems evaluation generally measures the quality of the algorithm, or system, *offline*, i.e. based on some information retrieval metric, e.g. precision or recall. The metrics do however not always reflect the users' perceptions of the recommendations. Perception-related values are instead often measured through user studies, however the bulk of the work on recommender systems is evaluated through offline analysis. In the work presented in this paper we choose to neglect the quality of the recommender system and instead focus on the similarity of aspects related to users' perception of recommender systems. Based on a user study ($N = 132$) we show the correlation of concepts such as *usefulness*, *ratings*, *obviousness*, and *serendipity* from the users' perspectives.

## Categories and Subject Descriptors

H.3.4 [**Information Technology and Systems Applications**]: Decision support; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - Information filtering, Retrieval models, Selection process

## General Terms

Algorithms, Design, Experimentation, Human Factors, Measurement, Performance

## Keywords

Recommender systems, analysis, personalization, user-centric evaluation

## 1. INTRODUCTION

During the last decade, techniques for item-centric recommendation have led to steady improvements in the accuracy of recommender systems. If we suppose that the goal of a recommender system is to accurately predict ratings, this improvement should be considered as progress. However, it is worth questioning whether this progress actually leads to better recommendations - or whether we are missing a broader point? Today, recommender systems are understood through the way they are evaluated, in most cases this maps to some measure of predictive accuracy. Without measuring user satisfaction, there is no means to map the common measures (e.g. predictive accuracy) back to the users of an evaluated system [6]. User studies relating to recommender systems often estimate the perceived quality of a system based on a set of questions asked to users either during the interaction with the system, or at a later point in time. In this paper we ask the question: "How do user-centric aspects of recommender system evaluation map onto each other?"

The most direct way to answer this is by performing user studies. Specifically, with the goal in mind of examining the similarity of common user-centric evaluation aspects, this paper conducts a comparative user study using three distinct recommendation approaches, and performs a similarity analysis of said aspects. The first of the three models used for this purpose is $k$-nearest neighbors (kNN), where users are clustered into neighborhoods, based on item rating history, and these neighborhoods are used as a basis to generate future recommendations. kNN is arguably one of the most commonly used algorithms in item-based recommender systems and has been well studied, e.g. [2]. The second model is a $k$-furthest neighbors model (kFN) [9], which essentially turns kNN inside out by first creating neighborhoods based on the least similar users in terms of ratings, followed by recommending items disliked by the neighborhood. The third model is a random recommender, not taking any information about users, or items they have interacted with, into consideration.

By using these three diverse approaches (in terms of methods of recommendation and quality measured in traditional performance metrics), our assumption is that answers given in the user study reflect a non-algorithm-related state of users' perception of evaluation concepts.

## 2. RELATED WORK

When considering metrics for evaluation of recommender systems, focusing on the user and on user satisfaction is vital. However, a number of issues are raised when people are brought into the equation, making the evaluation process more complicated.

Recently, a large amount of research has been focusing on user-centric aspects of recommender system evaluation, e.g. [1, 3, 7, 8] to mention just a few. Knijnenburg et al. [7] presented a pragmatic approach to user-centric evaluation

*IIiX* 2012, Nijmegen, The Netherlands
Copyright 2012 ACM 978-1-4503-1282-0/2012/08 ...$15.00.

of recommender systems, formalizing some of the aspects involved. Bollen et al. [1] evaluated how the number of available choices affects the perceived quality of a recommendation from the users' perspective.

McNee et al. [8] approached the concept from a different perspective, evaluating how better performance in terms of accuracy metrics could be detrimental to the overall perceived quality of a system, like many related works they recommend recommender systems researchers and developers to employ user-centric evaluation techniques.

Additional research directions undertaken focus on how user interfaces can be used to increase (or decrease) the perceived quality of recommender systems, e.g. Hu and Pu [5].

# 3. EXPERIMENTS

In order to examine the relationship between different evaluations of common quality criteria in recommendation systems, an online user study was created and users were invited to participate. The study was implemented in the form of a simple movie recommendation system, participants were asked to answer a set of questions upon having been given a set of recommendations.

## 3.1 Movie Recommendation Study

The user study consisted of two steps, first participants were asked to rate a minimum of 10 movies from a page showing a random selection of 100 of 500 most rated movies in the Movielens 10 million dataset[1], shown in Fig. 1.

Having rated at least 10 movies, the system generated 10 recommendations based on the Movielens rating dataset. For the sake of comparison, the participating users were presented with recommendations generated either by one of the neighborhood models (kNN or kFN), as described in Section 1, taking their ratings into consideration, or by a random recommender not taking the ratings from the previous step. Each user was presented with a page containing the top 10 recommendations and a set of questions, see Fig. 2.

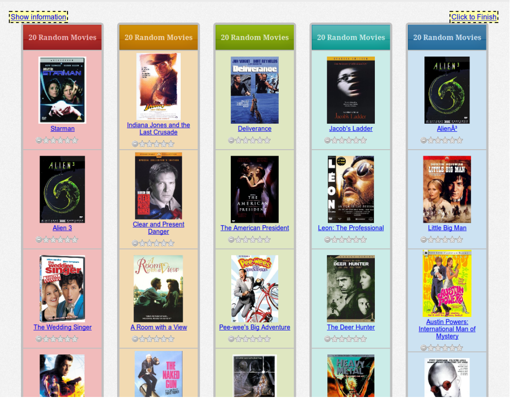
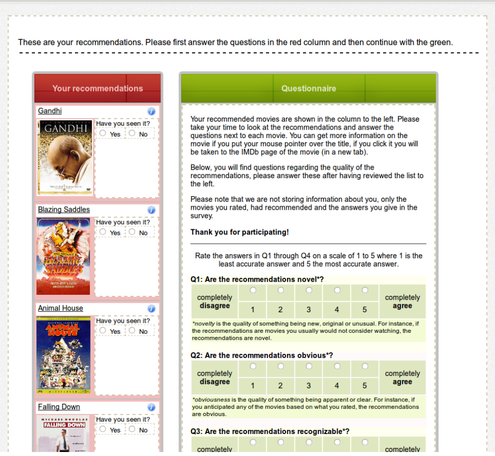


**Figure 1: The movie rating page where participants were asked to rate movies. The page contains a random selection of 100 of the 500 most popular movies in the Movielens 10 million dataset.**



These are your recommendations. Please first answer the questions in the red column and then continue with the green.

Your recommendations

Gandhi

Have you seen it? ○ Yes ○ No

Blazing Saddles

Have you seen it? ○ Yes ○ No

Animal House

Have you seen it? ○ Yes ○ No

Falling Down

Have you seen it? ○ Yes ○ No

Questionnaire

Your recommended movies are shown in the column to the left. Please take your time to look at the recommendations and answer the questions next to each movie. You can get more information on the movie if you put your mouse pointer over the title, if you click it you will be taken to the IMDb page of the movie (in a new tab).

Below, you will find questions regarding the quality of the recommendations, please answer these after having reviewed the list to the left.

Please note that we are not storing information about you, only the movies you rated, had recommended and the answers you give in the survey.

**Thank you for participating!**

Rate the answers in Q1 through Q4 on a scale of 1 to 5 where 1 is the least accurate answer and 5 the most accurate answer.

**Q1: Are the recommendations novel*?**

completely **disagree** ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 completely **agree**

**novelty* is the quality of something being new, original or unusual. For instance, if the recommendations are movies you usually would not consider watching, the recommendations are novel.

**Q2: Are the recommendations obvious*?**

completely **disagree** ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 completely **agree**

**obviousness* is the quality of something being apparent or clear. For instance, if you anticipated any of the movies based on what you rated, the recommendations are obvious.

**Q3: Are the recommendations recognizable*?**

completely ○ ○ ○ ○ ○ completely

**Figure 2: The questions presented to participants in the survey after having rated a minimum of 10 ratings.**

As opposed to both neighborhood-based methods, the computation time of the random recommender is negligible. To create the illusion of generating personalized recommendations, one of the neighborhood-based algorithms was trained (but not used) in parallel to serve as a timer.

### *Questionnaire*

For each of the 10 recommended movies (shown in the leftmost column in Fig. 2), participants were asked to answer a set of questions. Questions were chosen to reflect standard quality measuring aspects in recommender systems (e.g. usefulness, recognizability, customer retention) as well as reflecting the current state of the art in recommender system quality measurement (e.g. serendipity, novelty) [6]. The first set of questions (relating to each of the 10 recommended movies) was

**Have you seen the movie?**

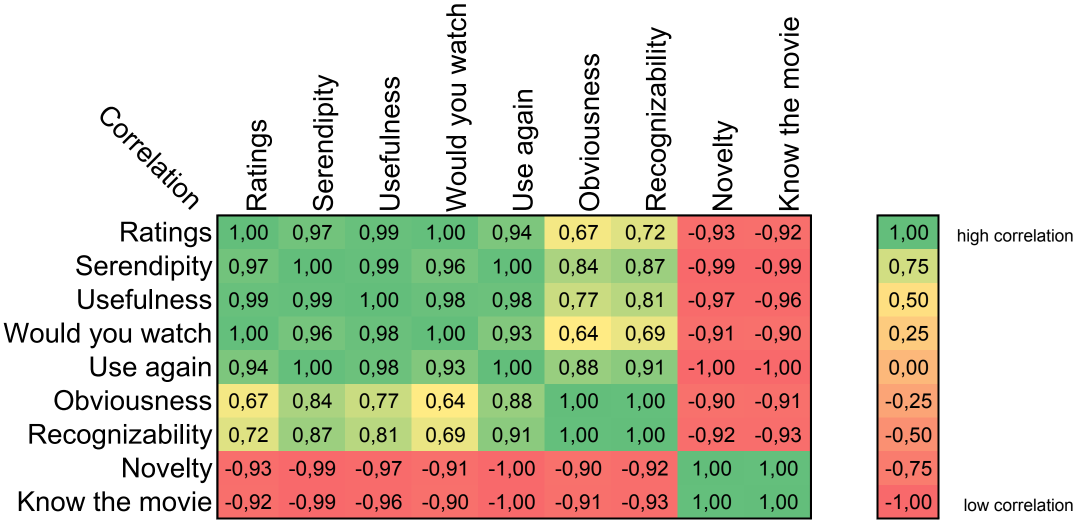
if Yes: Please rate it (5 star rating)

if No:
I. Are you familiar with it? (y/n)
II. Would you watch it? (y/n)

Additionally, participants were asked to answer 8 questions regarding the complete set of recommended items. The questions were:

1. Are the recommendations novel?
2. Are the recommendations obvious?
3. Are the recommendations recognizable?
4. Are the recommendations serendipitous?
5. Are the recommendations useful?
6. Pick the movie you consider the best recommendation.
7. Pick the movie you consider the worst recommendation.
8. Would you use this recommender again?

[1] http://www.grouplens.org/system/files/ml-10m-README.html



| Correlation | Ratings | Serendipity | Usefulness | Would you watch | Use again | Obviousness | Recognizability | Novelty | Know the movie |
|---|---|---|---|---|---|---|---|---|---|
| Ratings | 1,00 | 0,97 | 0,99 | 1,00 | 0,94 | 0,67 | 0,72 | -0,93 | -0,92 |
| Serendipity | 0,97 | 1,00 | 0,99 | 0,96 | 1,00 | 0,84 | 0,87 | -0,99 | -0,99 |
| Usefulness | 0,99 | 0,99 | 1,00 | 0,98 | 0,98 | 0,77 | 0,81 | -0,97 | -0,96 |
| Would you watch | 1,00 | 0,96 | 0,98 | 1,00 | 0,93 | 0,64 | 0,69 | -0,91 | -0,90 |
| Use again | 0,94 | 1,00 | 0,98 | 0,93 | 1,00 | 0,88 | 0,91 | -1,00 | -1,00 |
| Obviousness | 0,67 | 0,84 | 0,77 | 0,64 | 0,88 | 1,00 | 1,00 | -0,90 | -0,91 |
| Recognizability | 0,72 | 0,87 | 0,81 | 0,69 | 0,91 | 1,00 | 1,00 | -0,92 | -0,93 |
| Novelty | -0,93 | -0,99 | -0,97 | -0,91 | -1,00 | -0,90 | -0,92 | 1,00 | 1,00 |
| Know the movie | -0,92 | -0,99 | -0,96 | -0,90 | -1,00 | -0,91 | -0,93 | 1,00 | 1,00 |

| | |
|---|---|
| 1,00 | high correlation |
| 0,75 | |
| 0,50 | |
| 0,25 | |
| 0,00 | |
| -0,25 | |
| -0,50 | |
| -0,75 | |
| -1,00 | low correlation |

**Figure 3: The Pearson Product-Moment Correlation of the answers given to questions across all three recommendation algorithms. Each row/column corresponds to the correlation of the questions asked in the questionnaire.**

Participants were asked to answer the first five of the latter questions stating the level of agreement, from strongly disagree (1) to strongly agree (5). A short description was given for the terms *novel*, *obvious*, *recognizable* and *serendipitous* in order to mitigate erroneous answers based on misunderstanding of the question (see Appendix A). Additionally a field for comments was placed on the bottom of the survey.

Submission of the answers was only possible after all questions had been answered.

The questionnaire was designed to create a truthful picture of users' perceived usefulness of the recommendations, considering aspects such as choice overload [1], construction of questions [6] and similar perception-related concepts [4, 10].

## 3.2 Data

The link to the survey was circulated on social media sites (e.g. Facebook, Twitter, Google+) during the last two weeks of March 2012. People from several professions (computer science professionals, lawyers, business management professionals) were asked to circulate the link in their networks in order to gain participants from several communities, so to mitigate biasing effects which could surface in answers from a homogeneous community. The data used in this paper was collected in early April 2012, a total of 132 participants had completed the survey at the time.

Knijnenburg et al. claim that "at least 20 users" per condition should be adequate for being able to mine statistically sound data from a user study [7], indicating the amount of participants in our study should be sufficient.

No demographic data except for location (reverse lookup via IP address) was collected. The participants of the survey came from a total of 19 countries on 4 continents. Participants from Germany, Ireland and Sweden were most prominent, in descending order.

In order to analyze the similarity of the questions, the answers from all users were averaged on a per-question per-algorithm basis creating a vector containing three values, the average value for the question across the three recommenders. The question vectors were then compared using the Pearson Product-Moment Correlation Coefficient (Eq. (1)) towards each other according to

$$correl = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} \qquad (1)$$

where $X$ and $Y$ are the average values per question per algorithm. The resulting similarities are shown in Fig. 3.

# 4. RESULTS & DISCUSSION

Fig. 3 summarizes the pairwise similarities between the questions. The most dissimilar questions towards other questions are those regarding *novelty* and *recognizability.* They are however very similar towards each other, which is perhaps expected as they both, essentially, answer the same questions; one for unrated movies, and the other for rated movies (recall the questionnaire layout in Fig. 2).

Questions regarding *serendipity*, *usefulness*, whether the users would *use the system again* and the *ratings* seem to belong to a cluster of questions being perceived similarly. The implication of this being that the questions seem to correspond to similar values, e.g. a serendipitous movie will often be rated highly, or if a user considers a system useful the probability of using it again is higher, etc.

As for the less similar questions, i.e. ratings and willingness to watch a certain movies vs. obviousness and recognizability, the similarities tell us these questions are not entirely unrelated, however considerably less than, for instance, ratings and usefulness. We believe this is due to aspects such as obviousness and recognizability belong to the same concepts, i.e. once again how obvious a set of recommendations is probably related to how many movies are recognized, as something which is unknown can still be obvious if the initial step (the ratings performed in Fig. 1) only contains unknown items, and vice versa.

The main observations are: high ratings correlate with questions considering serendipity, usefulness and retention (use again, watch the movie). Obviousness and recognizability tend be of little importance, but do not seem to discourage people from using the system again. High levels of unknown/novel movies tend to point to lower ratings and less consumption (watching the movie).

It should however be noted that all observations are averaged with respect to the algorithms, expanding the study to include more algorithms diverse recommendation algorithms could alter the similarities.

## 5. CONCLUSION & FUTURE WORK

In this work we have presented how certain common aspects used for user-centric evaluation of recommender systems are perceived by users, in relation to each other. Our analysis points to high similarity between the concepts of serendipity, usefulness, willingness to consume, willingness to use a recommendation system again and ratings. Whereas concepts such as recognizability and novelty seem to have a similarity towards each other (i.e. complementing each other), these concepts however are completely unrelated to the aforementioned aspects.

Due to the currently immature analysis of our results, we do not present claims towards which questions should be used in comparison and evaluation of recommender system. In Fig. 3 we present an overview of similarity levels between the questions we evaluated, showing high and low similarity between all evaluated questions. The figure can be used as a rule of thumb when creating user studies for the purpose of evaluating a recommender system.

In real-world applications, we often only have access to users' ratings. With this in mind, we can infer that the ratings may also hold information about the perceived quality of the recommender in terms of serendipity, usefulness, intention to watch movies, and intention to use the system again.

We are currently in the process of performing an exhaustive analysis of the results presented in this work; this analysis will serve as the basis for continued research in the context of user-centric evaluation, perception-oriented aspects, and other related concept in recommender systems.

## 6. ACKNOWLEDGMENTS

The authors would like to thank all participants of the study, everyone who helped by distributing it in their communities, Michael Meder at TU Berlin for help with the implementation of the user study, and Benjamin Fields at Musicmetric and Bart Knijnenburg at UC Irvine for their help and feedback throughout the study.

## 7. REFERENCES

[1] Dirk Bollen, Bart P. Knijnenburg, Martijn C. Willemsen, and Mark Graus, 'Understanding choice overload in recommender systems', in *Proc. of the 4th ACM conference on Recommender systems*, RecSys '10, pp. 63–70, New York, NY, USA, (2010). ACM.

[2] Laurent Candillier, Kris Jack, F Fessant, and Frank Meyer, 'State-of-the-art recommender systems', *Collaborative and Social Information Retrieval and AccessTechniques for Improved User Modeling*, 1–22, (2009).

[3] Dan Cosley, Shyong K. Lam, Istvan Albert, Joseph A. Konstan, and John Riedl, 'Is seeing believing?: how recommender system interfaces affect users' opinions', in *Proc. of the SIGCHI conference on Human factors in computing systems*, CHI '03, pp. 585–592, New York, NY, USA, (2003). ACM.

[4] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl, 'Evaluating collaborative filtering recommender systems', *Trans. Inf. Syst.*, **22**(1), (2004).

[5] Rong Hu and Pearl Pu, 'Enhancing recommendation diversity with organization interfaces', in *Proc. of the 16th Intl. conference on Intelligent user interfaces*, IUI '11, pp. 347–350, New York, NY, USA, (2011). ACM.

[6] Bart Knijnenburg, Martijn Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell, 'Explaining the user experience of recommender systems', *User Modeling and User-Adapted Interaction*, **22**, 441–504, (2012).

[7] Bart P. Knijnenburg, Martijn C. Willemsen, and Alfred Kobsa, 'A pragmatic procedure to support the user-centric evaluation of recommender systems', in *Proc. of the 5th ACM conference on Recommender systems*, RecSys '11, pp. 321–324, New York, NY, USA, (2011). ACM.

[8] Sean M. McNee, John Riedl, and Joseph A. Konstan, 'Solving the apparent diversity-accuracy dilemma of recommender systems', number 10, pp. 4511–4515, New York, NY, USA, (1998).

[9] Alan Said, Benjamin Kille, Brijnesh J. Jain, and Sahin Albayrak, 'Increasing diversity through furthest neighbor-based recommendation', in *Proc. of the WSDM'12 Workshop on Diversity in Document Retrieval (DDR'12)*, (2012).

[10] Kirsten Swearingen and Rashmi Sinha, 'Beyond Algorithms: An HCI Perspective on Recommender Systems', in *ACM SIGIR Workshop on Recommender Systems*, (2001).

# APPENDIX

## A. TERM DESCRIPTIONS

The following explanations were given to nontrivial terms used in the study. The explanations are based on explanations given on online dictionaries (e.g. The Free Dictionary[2]).

*novelty* the quality of something being new, original or unusual. For instance, if the recommendations are movies you usually would not consider watching, the recommendations are novel.

*obvious* the quality of something being apparent or clear. For instance, if you anticipated any of the movies based on what you rated, the recommendations are obvious.

*recognizable* in terms of how well known the recommendations are to you.

*serendipity* a "happy accident" or "pleasant surprise"; specifically, the accident of finding something good or useful without looking for it.

[2] http://www.thefreedictionary.com