Proceedings of the ACM RecSys 2010 Workshop on User-Centric Evaluation of Recommender Systems and Their Interfaces (UCERSTI),
Barcelona, Spain, Sep 30, 2010
Published by CEUR-WS.org, ISSN 1613-0073, online ceur-ws.org/Vol-612/paper4.pdf

User-perceived recommendation quality - factoring in the user interface

Mouzhi Ge
TU Dortmund
44221, Dortmund, Germany
mouzhi.ge@tu-dortmund.de

Carla Delgado-Battenfeld
TU Dortmund
44221, Dortmund Germany
carla.delgado@tu-dortmund.de

Dietmar Jannach
TU Dortmund
44221, Dortmund Germany
dietmar.jannach@tu-dortmund.de

ABSTRACT

Most works in the domain of recommender systems focus on providing accurate recommendations. However many recent works have raised the issue that beyond accuracy other aspects such as diversity and novelty also impact the quality of recommendations and the user/customer behavior. This initiative has opened up a new perspective regarding evaluating and improving recommendation techniques, but some challenges are still to be faced. For example, traditional evaluations of recommenders do not take into account the system's interface. While accuracy is a metric somehow uncoupled to the recommenders' interface, other metrics such as diversity and novelty are directly related to it: a user might better perceive a higher degree of diversity and novelty if this is emphasized by its interface. In this paper we discuss the relations between evaluation metrics, the recommender interface and the user-perceived recommendation quality. We present a general guideline to evaluate recommenders from perspectives other than accuracy and propose a general experiment design to investigate the effects of quality factors on recommendations taking into account the system's interface. We also show how the proposed experiment model could be used to experiment with the factors "diversity" and "novelty" and specifically show how these factors can be meaningfully introduced in an experiment. We believe that our current work can be used in future research as a basis example on how to exam the effects of evaluation metrics and the user interface in recommender systems.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Measurement techniques.

General Terms

Measurement, Performance, Reliability.

Keywords

Recommender system, experiment design, evaluation metric, diversity, novelty.

1. INTRODUCTION

The main goal of recommender systems is to provide personalized recommendations in order to improve users' satisfaction and assist the users in making decisions. Different recommender systems were developed and used in several domains over the last decades [1] and a variety of recommendation techniques were proposed. Accordingly, various metrics have been proposed to estimate the effectiveness and value of the recommender systems.

Several among the successful recommendation techniques are based on a prediction of the degree to which a user might like an item. Because of this, the traditional evaluation approaches for recommenders are focused on the accuracy of the generated predictions, based for example on the Mean Absolute Error. Such approaches focus on the algorithm used to generate the recommendations, but do not look at the system as a whole. Usually these measurements are done in offline experiments [12] that do not take into account the user interaction with the system. Thus, such evaluations are typically independent of the system's interface and uncoupled from the user experience.

Although it is clear that the accuracy of the recommendations can affect the perceived quality of the system and the customer/user behavior, recent works argue that there are other important aspects we need to take into account [8, 14]. Several aspects of the perceived value of a recommender depend on the user interface and cannot be captured in an offline-experimental setting, in which e.g. only the ratings are available. According to Francisco Martin, who was RecSys09 keynote, up to 50% of the value of recommenders comes from a well-designed interface. Although this hypothesis is not supported by empirical evidence yet, we indeed believe that the interface of a recommender has a strong effect on its perceived value, and also that changes on the interface will affect the user's perception of the recommendations.

A classical example of the impact of the interface in the user perception of the recommendations is the case of serendipitous items in recommendation lists. When implemented inappropriately, unexpected items in the recommendation list may leave the user with the impression that the system does not understand his real needs, and therefore he may stop following the recommendations or even stop using the system. These risks can be reduced by the use of more (visual) explanations/clues regarding the reasons as to why an item was recommended. This has been done by Amazon.com (*http://www.amazon.com*), where one can see different lists of items classified by headers like "Users that bought this also bought that", "your recommendations" and "special offers" [16]. In this manner, the risk of misinterpreting the principle of the recommendation is reduced.

Several authors have already discussed quality factors beyond accuracy that may influence recommendations, e.g. [9, 13] and also how to use these factors to evaluate recommendations [8, 14, 15, 17]. In particular, [7] touched the matter of the advantages of online over offline evaluation strategies. We consider this a very important step towards exploiting the possibilities that different quality factors can bring to recommenders, but at the same time we believe there is a second step to be made: incorporating the interface and user interaction in the evaluations. Indeed, reports on experiments where quality factors were analyzed together with the recommenders interface already appeared on the literature [4, 20]. In this paper we approach this topic directly and discuss a

Proceedings of the ACM RecSys 2010 Workshop on User-Centric Evaluation of Recommender Systems and Their Interfaces (UCERSTI),
Barcelona, Spain, Sep 30, 2010
Published by CEUR-WS.org, ISSN 1613-0073, online ceur-ws.org/Vol-612/paper4.pdf

general evaluation approach that incorporates the system interface. Our main point is that there is a strong influence of the interface on the user perception of the quality of the recommendations received, and experiments that neglect this influence may lead to biased conclusions.

The paper is organized as follows. In Section 2 we describe our general model for representing the relationship among recommendation quality factors, user interface and customer behavior. In Section 3 we give an example of how the model can be instantiated into a specific experiment design. Section 4 focuses on how to incorporate the quality factors "novelty" and "diversity" in an experiment. Section 5 presents our conclusions and plans for future work.

2. MODEL CONSTRUCTION

As mentioned above, we argue that the user perception of different recommendation quality factors may be significantly affected by the system interface. Generally, different user interfaces can be used to present the recommendations. Therefore the user interface can be considered a moderator variable that affects the direction and/or strength of the relation between the recommendation quality factors and the customer behavior.

We propose a general model to examine the relationship among recommendation quality factors, user interface and customer behavior. The model is described as follows.

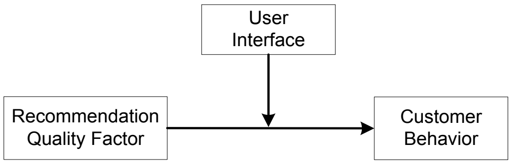

Figure 1: General model of the relationship among recommendation quality factors, user interface and customer behavior.

In our model, "recommendation quality factor" is a general term that represents the several possible factors that indicate different quality aspects of the recommendations. A few factors have been proposed in previous research such as for example diversity, novelty, serendipity and coverage [8, 14]. Also, "user interface" is considered in the context of recommender systems as a display format that allows the customers to interactively explore the recommendations. For example, a recommendation can be visually represented using plain text or a picture (as indicated by [10]). By "customer behavior" we mean the customers' actions or responses that may be affected by recommenders such as customer purchase behavior [2], customer decision making [18], customer interests [19], or satisfaction [20].

The more quality factors we include, the more different interfaces might be used to express the recommendations with different effects on the user perception (it is always the case that different interfaces can be used, but if no quality factors are added there might not be any effects on the user perception). The goal of our model is to analyze the interactions between user interface, quality factors and customer behavior.

When instantiating the model, we are still facing the following questions: How to measure the recommendation quality factors? Which interface can be used to express the recommendations? How to measure customer behavior? In the next section, we develop a first research design of how to implement our model.

3. MODEL INSTANCE AND EXPERIMENT DESIGN

It has been found that experimental research is an effective approach to address cause and effect relationships [3, 11]. In order to show how our general model can be instantiated within a concrete experiment, we selected two well cited evaluation metrics as recommendation quality factors: *diversity* and *novelty,* and use two common interface styles to visualize the recommendations: *single list* and *multiple lists*. The customer behavior is analyzed in terms of *purchase rate* and *customer satisfaction*, since these are the typical indicators for recommender's performance.

Thus, two independent variables are determined, each of which has two possible values: diversity (with or without), novelty (with or without). The values for the variable user interface are also two: single list or multiple lists. The customer behavior is determined by two dependent variables: customer purchase and customer satisfaction. On one hand, the customer purchase is mainly the vendor's perspective and aims at directing customers to adopt or buy the recommended product regardless of their satisfaction. It can be directly measured by the sales increase generated from the recommender system. On the other hand, customer satisfaction stands for the customer perspective and how the recommended products or the recommendation session as a whole fulfilled his expectations. It is usually measured by a survey using Likert scales.

Figure 2 gives an overview of the instantiated model, once the independent and dependent variables are determined.

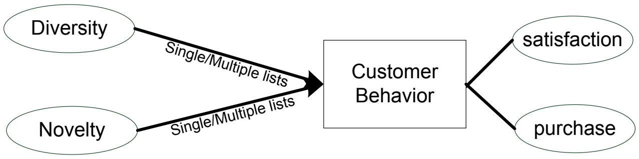

Figure 2: Instantiated model with dependent and independent variables.

To exemplify the usage of our model we designed the following experiment:

"A movie website with recommendations is presented to the experiment participants. First, the participants are asked to register and enter their movie preference. After that, each participant will obtain 3 *electronic vouchers* that can be use to buy 3 movies. Each voucher can be used to buy one movie. Next, the users are presented with a list of recommendations. Based on these recommendations, the participants will select movies for purchase. They can use all the vouchers at one time or save the vouchers for the future. After choosing the movies, we present a survey to the participants and ask if they are satisfied with the recommendation system."

We assume that the participants will carefully choose movies as they can postpone the choice and use the points for future movies in case they do not feel "tempted" by any of the recommended items. To present the recommendations, the user interface is also randomly selected. That means the recommendations can be presented only in a single list or in multiple lists that are used to separate basic, diverse and novel recommendations. In Table 1 we present a sketch of the factor design for the movie website experiment. The situations 1 to 4 in the table describe different configurations for the independent and moderator variables.

Table 1. Factor design for the movie website experiment.

	Diversity	Novelty	User interface	
			Single List	Multiple Lists
1	without	without	Basic	Basic lists
2	with	without	0.3 diversity	Basic list Diverse list (1.0 diversity)
3	without	with	0.3 novelty	Basic list Novel list (1.0 novelty)
4	with	with	0.3 diversity, 0.3 novelty	Basic list Diverse list (1.0 diversity) Novel list (1.0 novelty)

The basic list contains the items selected by the recommendation algorithm. Novel lists are generated by manipulating the basic recommendation lists to include more novel items among the *top-n* items, considering that *n* is the number of items that will be presented to the user. In situations 2 to 4 in Table 1, a list with values 0.3 novelty stands for one list with novelty degree of 0.3, and a similar approach is used for diverse lists. This will be further explained in Section 4.

Diversity and novelty of recommendations can be designed in the form of binary or continuous. While the binary form is used to define the recommendations with or without diversity and novelty, continuous form defines diversity and novelty in the sense of various percentages. With the binary design, the recommendations can be configured as a between-subjects factor design. Thus, in each interface the result can be analyzed accordingly using a two-way ANOVA analysis [12]. When we employ the continuous design, the result can be analyzed using a regression analysis. The two designs can mutually confirm or supplement their findings. In addition, the experimental result of the effect of different interface designs can be analyzed based on A/B split testing. Based on this data, we can then analyze if and to what extent diverse or novel recommendations affect customer behavior and how to provide an appropriate interface when we introduce more diversity and novelty into the recommendations.

4. EVALUATING NOVEL AND DIVERSE RECOMMENDATIONS

In this section we focus on two quality factors: novelty and diversity. Novelty is related to items the user was not aware of. Diversity is generally defined as the opposite of similarity [17]. To implement these quality factors in an experiment, we should be able to control the degree of novelty and diversity in a recommendation list and multiple lists, considering the specific user interfaces involved in the experiment.

We propose to use a combination of three factors to identify novel items in a list of items: (1) “freshness” of an item (i.e., items that were recently launched), (2) non-popularity (popular items are not considered novel) and (3) limited relation to the long-term user profile (e.g. by previous ratings, feedback or views of this item in previous sessions). Each item is scored according to each factor in a scale from 0 (low) to 1 (high), and then the degree of novelty of the item is calculated simply by the average of the score for the three factors. Considering that i is an item and that $0 \leq fresh(i), nonpop(i), unknw(i) \leq 1$ represent respectively the degree of freshness, non-popularity and lack of relation to the user long term profile, the novelty $nov(i)$ of item i is defined by:

$$nov(i) = \frac{fresh(i) + nonpop(i) + unknw(i)}{3}$$

To calculate the degree of novelty of a recommendation list, we use the novelty degree of the list items. Assuming that an item is considered to be novel if its novelty degree is higher, for example, 60%, the novelty degree of a recommendation list L is the proportion of items from the list whose novelty degree surpasses this threshold.

$$nov(L) = \frac{|\{i \in L \,|nov(i) > 0.6\}|}{|L|}$$

For the multiple lists interface, we propose two lists: the basic list (as directly given by the recommendation algorithm) and a list consisting only of novel items, $nov(L2) = 1.0$. Another option is to have one list with a low degree of novelty (e.g. $(L1) = 0.2$) and another with a high degree of novelty (e.g. $nov(L2) = 0.6$).

The most explored method for measuring diversity uses item-item similarity. The diversity of a list of items can then be measured based on the sum, average, minimum or maximum distance between pairs of items [17, 19]. We adopt a slight modification of the approach from [20] and use the intra-list similarity metric (ILS). Considering that B is a set of items, this metric is based on a function of $c: B \times B \rightarrow [0, 1]$ that is supposed to measure the similarity between two items according to a predefined criterion. Then we calculate the ILS as follows:

$$ILS(L) = \frac{\sum_{i_k \in L} \sum_{i_e \in L, i_k \neq i_e} c(i_k, i_e)}{2}$$

The selection of function c is dependent on the available content information for each item and can also be dependent on the user’s preferences. The simplest option is to consider c as the degree of intersection of the items’ properties (such as size, color, weight or genre). If we consider a function $prop: B \times P \rightarrow \{0, 1\}$ where P is the set of available item properties, we can define c as:

$$c\ (i_k, i_e) := \frac{\sum_{p \in P} prop(i_k, p).\, prop(i_e, p)}{|P|}$$

We therefore can measure the diversity of a list by means of ILS and c. As high values of ILS denote low diversity, we take the inverse of ILS and define $DIV(L)$ as the degree of similarity of a list of items.

$$DIV(L) = \frac{1}{ILS(L)}$$

When designing diversity in multiple recommendation lists, we can use a threshold to discriminate a list with high similarity (e.g. $DIV(L) > 0.6$) or one with low similarity (e.g. $DIV(L) < 0.3$).

The measurements $nov(L)$ and $DIV(L)$ can be manipulated by changing some of the items in L. One strategy to increase $nov(L)$ is to substitute some items j from L by an equal number of items i not yet in L, such that $nov(i) > 0.6 > nov(j)$; considering that the threshold used to calculate $nov(L)$ is 0.6. In a similar way we can manipulate the diversity of a list L by replacing items that differ very little from other items already in L, i.e., items i for which $c(i, k)$ is high for several other elements k from the same

Proceedings of the ACM RecSys 2010 Workshop on User-Centric Evaluation of Recommender Systems and Their Interfaces (UCERSTI),
Barcelona, Spain, Sep 30, 2010
Published by CEUR-WS.org, ISSN 1613-0073, online ceur-ws.org/Vol-612/paper4.pdf

list L should be replaced by other items j, not yet in L, for which $c(j,k)$ is low for as many items k from L as possible.

5. CONCLUSION

Although many previous evaluations of recommender systems used accuracy as the only quality factor to be taken into account, recent works have shown that other metrics are also related to the user perception of quality of the recommendations. A further investigation of quality factors as diversity, novelty and serendipity lead us to conclude that the users' perception of these factors is highly linked to the system's interface.

This paper modeled the relations between evaluation metrics, the recommender interface and the customer behavior. Our main contribution is a first general model for experiments to evaluate recommender systems that aim to investigate the effects of recommendation quality factors and user interface on customer behavior. Besides proposing this general model, we also provided an example of how to instantiate the model for one specific experiment concerned with the evaluation of diversity, novelty and interface on customer purchase and satisfaction.

We consider that exploring other quality factors than accuracy is an important step towards improving the impact of recommenders and strongly believe that factoring in the user interface is crucial to realistically evaluate the users' perception of the quality of the recommendations. An interesting point is that the expected results could be used in line with the business strategy. For example, presume we found that with multi-list interface, diversity significantly affects customer purchase. Thus if we intend to promote certain product, we have more chances to advertise this product by including it in the diverse recommendations.

Although the designed experiment has not yet been completed, our study provided a first guideline on how to incorporate user interface aspects better in the recommender systems' evaluation. The experimental results are expected to show how the quality of recommendations could be optimized by presenting appropriate user interfaces. We intend to contribute to future research on how to examine the effects of evaluation metrics and the user interfaces in recommender systems so that further quality factors can be meaningfully evaluated.

6. REFERENCES

[1] Adomavicius, G. and Tuzhilin, A. 2005. Towards the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6), pp. 734-749.

[2] Bodapati, A.V., Recommendation systems with purchase data. Journal of Marketing Research. 45(1). pp. 77-93.

[3] Campbell, D.T. and Stanley, J. (1963), Experimental and quasi-experimental designs for research. Houghton-Mifflin, Boston, Massachusetts, USA.

[4] Chen, L. and Pu, P. 2007. Preference-Based Organization Interfaces: Aiding User Critiques in Recommender Systems. Lecture Notes In Artificial Intelligence, vol. 4511. pp 77-86.

[5] Good, N., Schafer, J., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J. and Riedl, J. 1999. Combining collaborative filtering with personal agents for better recommendations. Conference of the American Association of Artificial Intelligence, Florida, USA. pp. 439-446.

[6] Gronroos, C. 1983. Strategic management and marketing in the service sector. Marketing Science Institute. USA.

[7] Hayes, C. Massa, P., Avesani, P., and Cunningham, P. 2002. An on-Line Evaluation Framework for Recommender Systems. Workshop on Personalization and Recommendation in E-Commerce, Malaga, Spain.

[8] Herlocker J., Konstan J., Terveen L. and Riedl J. 2004. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22(1), pp. 5–53.

[9] Iaquinta, L., Gemmis, M., Lops, P. and Semeraro, G. 2008. Introducing serendipity in a content-based recommender system. 8th International Conference on Hybrid Intelligent Systems, Barcelona, Spain. pp 168-174.

[10] Jannach, D., Hegelich K.: 2009. A case study on the effectiveness of recommendations in the Mobile Internet, ACM Conference on Recommender Systems, New York, pp. 205-208.

[11] Jarvenpaa, S.L., Dickson, G.W. and DeSanctis, G. 1985 Methodological issues in experimental IS research: experiences and recommendations, MIS Quarterly, 9(2), pp.141-156.

[12] Juran, J.M., Gryna, F.M. and Bingham, R.S. 1974. Quality control handbook, 3rd edition, McGraw-Hill, New York, USA.

[13] Kamahara, J., Asakawa, T., Shimojo, S. and Miyahara, H. 2005. A community-based recommendation system to reveal unexpected interests. 11th International Multimedia Modeling Conference, Melbourne, Australia. pp. 433 – 438.

[14] Mcnee, S., Riedl, J and Konstan, J. 2006. Accurate is not always good: How Accuracy metrics have hurt recommender systems, Conference on Human Factors in Computing Systems, Quebec, Canada. pp. 1-5.

[15] Murakami, T., Mori, K. and Orihara, R. 2008. Metrics for evaluating the serendipity of recommendation lists. New frontiers in artificial intelligence, Lecture Notes in Computer Science, vol. 4914 pp. 40-46.

[16] Schafer, B., Konstan, J. and Riedl, J. 2001. E-Commerce Recommendation Applications, Journal of Data Mining Knowledge Discovery, 5(1-2), pp. 115-153.

[17] Shani, G. and Gunawardana, A. 2009. Evaluating Recommendation Systems. Microsoft research, Technical report, No. MSR-TR-2009-159.

[18] Senecal, S. and Nantel, J. 2004. The influence of online product recommendations on consumers' online choices. Journal of Retailing. 80(2), pp. 159-169.

[19] Zanker, M.; Bricman, M.; Gordea, S.; Jannach, D.; and Jessenitschnig, M. 2006. Persuasive online-selling in quality and taste domains. In Proceedings of 7th Intl. Conference E-Commerce and Web Technologies, Krakow, Poland, pp. 51–60.

[20] Ziegler, C., McNee, S. M., Konstan, J. A., and Lausen, G. 2005. Improving recommendation lists through topic diversification. In Proceedings of the 14th International World Wide Web Conference, Chiba, Japan , pp. 22-32.