Investigating Confidence Displays for Top-N Recommendations

Guy Shani & Lior Rokach & Bracha Shapira &
Sarit Hadash & Moran Tangi
Information Systems Engineering, Ben Gurion University

Recommendation systems often compute fixed length lists of recommended items to users. Forcing the system to predict a fixed length list for each user may result in different confidence levels for the computed recommendations. Reporting the system's confidence in its predictions (the recommendation strength) can provide valuable information to users in making their decisions. In this paper we investigate several different displays of the system's confidence to users and conclude that some displays are easier to understand and are favored by most users. We continue to investigate the effect confidence has on users, in terms of their perception of the recommendation quality and the user experience with the system. Our studies show that it was not easier for users to identify relevant items when confidence is displayed. Still, users appreciated the displays, and trusted them when the relevance of items is difficult to establish.

Introduction

Perhaps the most common task of recommendation systems is to present users with recommended items. For example, CNN[1] displays below the current news story, lists of other recommended stories, and YouTube[2] presents alongside the displayed clip a list of other suggested clip.

When a system is requested to provide a fixed number N of recommendations for every query (e.g. page-view), it is likely that while the system might be able to provide N recommendations that fit some queries (strong recommendations). However, on cases where the system is unable to provide N appropriate recommendations the user is provided with recommendations that are less appropriate to the query. The user's experience on such cases may cause her to reduce her trust in the system.

One solution to this predicament is to avoid displaying recommendations when the system is not confident in their quality (i.e., weak recommendations). However, this solution might cause the system to present very short lists of recommended items or even not display recommendations at all. A user who notices the lack of recommendations may suspect a system malfunction which could further degrade her trust in the system. An additional concern regarding this solution is that it requires a decision to be made about the acceptable level of confidence in the recommendations, namely, adding another tunable parameter to be set by the system administrator or perhaps personalized for each individual user.

An alternative to the above solution is adding a display of the system confidence in its recommendations or in the recommendation strength. Providing a display of confidence alongside a recommendation can help the user to decide on her personalized level of comfort with the recommendation's confidence. This also allows for context-based decisions. For example, when the user truly cannot find items of interest, she may be willing to consider less confident recommendations, whereas when the user has already found some interesting items, she may not be willing to consider such recommendations.

One must not confuse the predicted rating of the system with confidence. The system can predict that a user will give a 3 star rating to a movie with 99% confidence, or that she will give 5 stars to the movie with 10% confidence. The system may have an average error of 0.5 star, which is a measurement of its accuracy, but not of the system's confidence of its predictions. In this paper we limit ourselves to the case of Top-N recommendation tasks and leave the discussion of confidence displays for rating prediction to future research.

In this paper we take a close look at confidence displays for recommendations. We begin with suggesting a number of alternative displays for confidence. We conducted a user study to measure how well users grasp confidence displays given the variations. We conclude that discrete or continuous confidence displays that are based on well-known displays from other domains, such as the bar chart and fuel gauge, are easier to understand.

We continue to study the effect of confidence displays on users. We conduct a second user study in which we present users with three types of recommendations of degrading quality and ask the users to evaluate their quality, both with and without confidence displays. We examine whether confidence displays make it easier for users to identify high quality recommendations, and whether users learn to rely on the confidence display rather than directly evaluating the quality themselves.

[1] www.cnn.com

[2] www.youtube.com

Background

In this section we discuss relevant previous research. We first discuss the definition of confidence, and then review previous studies on confidence displays. We conclude with a discussion about various methods for computing confidence.

Confidence in a recommendation

In Statistics, alongside an estimate of a parameter, one typically reports on the significance of the estimation, which is the level of confidence that one has over the reported estimation. Such confidence computations can be used in order to prune out estimates that do not pass a predefined confidence threshold. The notion of presenting this confidence estimate to users as additional information was suggested many decades ago (Herman, Ornstein, & Bahrick, 1964).

In the literature of data mining and machine learning, there are a few papers that examine how the notion of confidence correlates with various predictive performance measures that are used to evaluate the performance of supervised learning techniques. For example, Esposito, Malerba, and Semeraro (1997) have shown that the classification accuracy is positively correlated with the confidence of the classification. Katz, Shabtai, Rokach, and Ofek (2012) have shown recently that the notion of confidence can be used to identify hard-to-classify instances and propose alternative classification. Rokach, Naamani, and Shmilovici (2008) have used the notion of confidence to determine which instances should be actively acquired in order to improve the predictive performance of a classifier.

The recommendation systems field, and the collaborative filtering approach in particular, have borrowed many ideas from statistics. Such an example is to measure the correlation between users as the correlation between random variables through the Pearson correlation. As such, it is not surprising that presenting confidence estimates to users was done even in earlier systems, such as MovieCritic[3].

In this paper we define confidence as the system trust in its own predictions or recommendations, as defined in statistics. This is to be distinguished from the user trust in the system, which is also sometimes referred to as confidence.

We focus here on Top-*N* recommendation tasks where the system is required to present the user with a fixed-length list of recommended items, as opposed to the rating prediction task, where the system is required to present the user with a predicted rating for a given item. In the context of Top-*N* recommendations, the confidence of the system in a prediction can also be interpreted as the recommendation strength, i.e., when the system is confident that a certain item is adequate for the active user. We can call this a strong recommendation. When the system is unsure whether an item is appropriate for the active user, we can call this is a weak recommendation. In this paper we use the term confidence, but the term strength is also valid.

Another related term in Top-*N* recommendation is "relevance". We can say that the system attempts to recommend only relevant items. Thus, a low confidence can be associated with questionable relevance. Clearly, when the system is confident that an item is not-relevant, it will not provide such a recommendation. When the system is displaying a non-relevant item, it may be because it was not confident as to whether it was relevant or not. Later, in our experiment, we use this intuition in picking low relevance items when we wish to model low confidence of the system in its predictions.

Confidence should not be confused with the utility of the recommended item, be it the system or user utility. For example, some items may contain more information to the user, other items might bring more profit to the system, some items may be more relevant to the user's goals, and so forth. In this work we are interested in the application where all items have the same utility. For example, in the image tagging application of our user study, we assume that all tags are equally useful. This would not be the case if, for example, each image could have been tagged with only 3 tags. If this was the case, then color tags might be less informative than object tags. For example, saying that an image contains the colors "red", "green", and "yellow", may not be as informative as saying that it contains the objects "flower", "leaf", and "droplets". Displays which combine two values, such as confidence and utility, are expected to be more complex, and are hence left for future research.

In this paper we focus on the confidence of a specific result returned by an algorithm. In some cases a confidence can be associated with an algorithm or a classifier in order to choose the best algorithm or to combine the results of several algorithms (e.g. (Dredze, Crammer, & Pereira, 2008; Littman, Keim, & Shazeer, 2002)). In the context of recommending items to users, it is difficult to see how such confidence scores may be presented to the user.

Confidence displays

Evaluating the presentation of recommendation confidence to users was first carried out within the MovieLens system. McNee, Lam, Guetzlaff, Konstan, and Riedl (2003) defined recommendations with low confidence as "risky" and displayed either a single die ("low risk") or two dice ("high risk") next to the recommended item. They evaluated their system in three scenarios with different level of risk-sensitivity context (risk-seeking, risk-neutral, risk-averse). In each scenario the participants were asked to choose a movie to watch. Subsequently McNee et al. (2003) evaluated several issues:

[3]MovieCritic.com was a website containing movie descriptions and opinions that is no longer active.

• Whether users notice the confidence displays, concluding that most users notice the dice, and after a minimal training, most understand the meaning of the confidence display.

• How users use the confidence display; finding that users that were already familiar with MovieLens (so called - "experienced users") clicked on risky items many times. This may indicate a bias in the user population of the study towards users who are interested in the study rather than in accomplishing their tasks.

• How happy are users with the MovieLens system; drawing a conclusion that confidence displays did not change the user attitude towards MovieLens.

• Whether training helps user; concluding that trained users perceive greater value in the confidence displays. Also, trained users perceived the system recommendations as more accurate when the confidence display was used.

• Whether users made use of the confidence displays, concluding that there is a correlation between the tasks' risk-sensitivity level and the way users approach items marked with the risk symbol (dice). In the risk-averse task, users avoided diced items. In the risk-seeking task, users looked for diced items, and for the risk-neutral task, users asked for information about the unsafe items but avoided selecting them.

• In this study subjects were experienced MovieLens users, familiar with recommendations for movies. It is reasonable to assume that these users have also viewed in the past the recommendations that were less risky through their standard MovieLens experience, and were thus more willing to try the more risky items.

Questions which were not directly evaluated in this study but are of interest to us are whether different graphical displays are easier or harder for users to understand and whether confidence displays help in estimating the quality of the recommendations. We also investigate whether users come to rely on confidence displays when they assess the system quality. Furthermore, McNee's study was conducted on a single domain, movies within MovieLens. In this paper we test whether the results also hold for other domains. Finally, the MovieLens system provides rating predictions while we study top *N* item recommendations. The difference between the two is that in rating prediction, the confidence display requires a presentation of two scores together (rating and confidence), which can be somewhat confusing. For item recommendations, the system only presents a list of recommended items which are typically with no score attached. As such, confidence displays become the only "score" presented to the user.

Confidence displays can also be considered as a part of the broader recommendation explanation area. Recommendation explanation research, in the hope of providing a better user experience (Tintarev & Masthoff, 2011; Herlocker, Konstan, & Riedl, 2000), suggests methods for providing additional information to users which can help them to better understand why certain items were recommended for them, increase trust in the system, and convince users to select the recommended items. In this respect, confidence displays could be seen as advancing the goals of the system, such as increased sales or increased user satisfaction. Estimating the effect of confidence displays over such goals is difficult without experimenting over a real recommender system with real users.

Computing confidence

Although the computation of confidence or strength estimates is not the focus of this paper, we note that such estimates are in some cases easy to obtain. In general, collaborative filtering algorithms become more confident as a user rates more items, or an item receives more ratings. We can thus associate a confidence measure for a specific rating prediction for an item with the number of ratings that the item received in our dataset (McNee et al., 2003). In the popular memory-based *k*-nearest neighbor methods (Desrosiers & Karypis, 2011) a confidence can be associated with, e.g. the average similarity of the *k* nearest users to the active user. In the more

Many classic classification and prediction algorithms from the machine learning literature can be augmented to provide a classification or prediction confidence estimation. For example, (Toth & Pataki, 2007) shows how a decision tree can out a "classification certainty" score; (Krzanowski et al., 2006) discusses various methods for computing the confidence in a specific classification result using a Bayesian approach. (Delany, Cunningham, Doyle, & Zamolotskikh, 2005) shows that naive Bayes classifiers and *k*- nearest neighbors techniques that are also popular in collaborative filtering applications do not provide calibrated confidence estimates.

There are some examples of collaborative filtering rating prediction algorithms (e.g. (Kadie, Meek, & Heckerman, 2002; Hofmann, 2003)) that take directly into account the uncertainty about the model prediction or the input data, when constructing statistical models. This is tightly related to the confidence of the algorithm in its prediction, and indeed with these methods a confidence measure is an immediate output of the algorithm. Karatzoglou and Weimer (2010) presents a version of the popular matrix factorization algorithm that can produce prediction confidence scores, in the form of a confidence interval.

When interested in item recommendations, as opposed to rating predictions, most systems compute a recommendation score for each item and then order the items by decreasing scores. For example, in Amazon (Linden, Smith, & York, 2003), the system computes an item-item Cosine correlation score. Another obvious method for ordering items is by decreasing conditional probability given the user profile or a specific item.

The conditional probability is interpreted as the probability that a user will like the item. When the system predicts that the user will like the item with a probability of 1, the system has complete confidence that the user will like the item. When the system predicts a probability of 0.5, the system is unsure whether the user will like the item or not. It is hence appropriate to think about the conditional probability as a confidence or strength measurement. For example, consider an item-item news story recommendation algorithm within an e-news website like `www.cnn.com` that uses conditional probabilities to recommend additional news stories. That is, given a viewer who is currently viewing news story i, the system will compute the conditional probability of reading any other story j by

$$pr(j|i) = \frac{count(i, j)}{count(i)} \quad (1)$$

where $count(i, j)$ is the number of users who viewed both story i and story j, and $cound(i)$ is the number of users who viewed item i. Thus, $pr(j|i)$ can be understood as the probability of viewing j after viewing i, and can also be understood as how confident the system is that a user that has viewed i will also view j.

A different view of confidence, which is more closely related to statistical significance we discussed above, is to consider the amount of evidence in favor of a prediction when computing confidence. For example, McNee et al. (2003) computed confidence through the amount of information over an item in the dataset. For them, items had a fixed "risk" score that was higher when there were only a handful of ratings for a particular item.

In this respect, confidence is tightly related to p-values; the probability of predicting a particular score by chance. As such, many methods for computing p-values from statistics could be used to compute confidence scores.

In many cases a similar strength or confidence measure can be estimated, but in general, this can be non-trivial. In this research we are mainly interested in the presentation of a confidence score, assuming that such a score can be computed. We hence leave the non-trivial exploration of methods for computing such scores to others.

Investigating Various Confidence Displays

We now begin our investigation of confidence displays for item recommendations. In this section we overview a set of confidence displays. We report the results of a user study that we have conducted to test whether users understand the confidence display.

We used the following principles (see, e.g. (Nielsen, 1994; Konstan, 2010)) in designing the displays below:

- Match between system and the real world: we selected many displays that map to well-known displays in the real world, such as the fuel gauge and the bar display, which is widely used to indicate, for example, cell-phone reception.
- Minimalist design: a variation in design from minimalistic designs, such as the up/down arrow to very rich displays, such as the confidence table. This allows us to explore whether a minimalistic design is indeed important in our setting.
- Consistency and standards: we maintained consistency with well-known standards, such as green indicating safety and red indicating risk (at least in the western-oriented cultures that the subjects in this study belong to), and an up arrow indicating a good status and a down arrow indicating a bad status.

All the displays below were adapted or designed by us for the purposes of this research. Some of them are naturally associated with recommendations, such as the stars interval and stars table which are commonly associated with movie ratings. Other displays can be associated with risk, like the slot machine, the stop light, the fuel gauge, and the up-down arrow. Other displays are naturally associated with strength like, for example, the bar chart, which is usually associated with cell-phone reception strength. Finally, some displays have a strong statistical association, such as the pie chart, the simple interval, and the confidence bars. We thus believe that these displays wrap up the spectrum of possible definitions and associations of confidence which we consider in this paper.

In this paper we are mainly interested in studying the effects of various confidence displays on users and not in designing the best possible confidence display. We are certain that other, perhaps more appropriate and useful displays can be designed and leave this for future research.

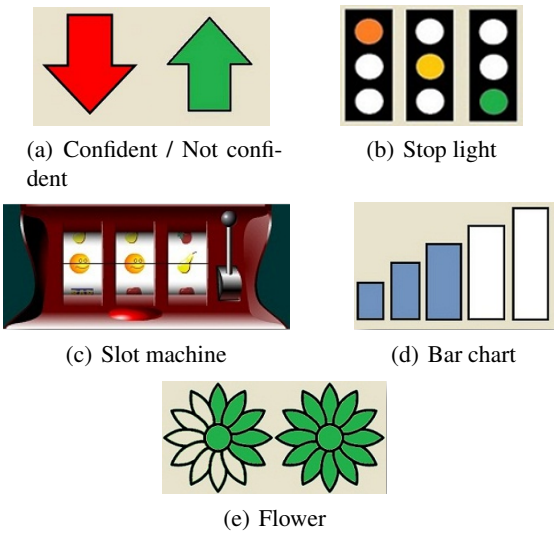
Discrete displays

Our first set of suggested displays focuses on discrete displays, i.e., displays that present a fixed number of confidence levels in various ways. Such displays may be easier to understand, as they do not require the interpretation of a numeric score and can be compared by the viewer quite easily. The displays vary in the number of available levels of confidence and the used graphic metaphor.

Up-down arrow: Figure 1(a) is the simpliest display — when the arrow is green and points up, the system is confident in its recommendation. When the arrow is red and points down, the system is not confident in its prediction.

Stop light: Figure 1(b) is a 3 level confidence level display. The green light is associated with confidence, the yellow light is associated with some risk, and the red light is associated with risky recommendations. This display associates confidence with risk and safety, and even somewhat prompts the notion that choosing items with low confidence is barred, like driving when the red light is on.

Slot machine: Figure 1(c) displays 4 levels of confidence.



(a) Confident / Not confident (b) Stop light (c) Slot machine (d) Bar chart (e) Flower

Figure 1. Discrete confidence displays.

This level of confidence is associated with the number of smiling faces that appear on the machine. The metaphor relates confident recommendations with luck or serendipity, i.e., a strong recommendation is a "lucky guess" by the system. This display is somewhat similar in motivation and representation power to the dice display employed by McNee et al. (McNee et al., 2003).

Bar chart: Figure 1(d) uses a 6 level confidence display (including the zero score where no bars are colored). Bars become colored from left to right and the level of confidence is denoted by the number of colored bars. This display is commonly used in cell phones to present the network signal strength and is thus familiar by most users.

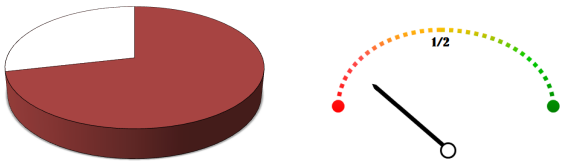
Flower: Figure 1(e) employs a 14 level confidence level (including the zero score where no petals are colored). The number of colored petals represents the confidence level. This is the confidence display with the highest number of discretization levels that we experimented with.

Continuous displays

The second set of images uses a continuous display to present the confidence estimation to the user. These graphical displays represent partial quantities. That is, displayed confidence is on a scale from 0 (not confident) to 1 (fully confident).

Pie chart: Figure 2(a) uses a pie chart to represent the level of confidence. As the pie becomes full, the associated level of confidence grows. This display is relatively well known by users and is commonly used in business analysis.

Fuel gauge: In Figure 2(b) The level of confidence is displayed by the angle of the needle in the gauge. As the needle leans to the right, the system is more confident, and as the needle leans to the left, the system is less confident. Furthermore, the green and red areas represent the safety and risky zones. This display again relates confidence with risk and safety.



(a) Pie chart (b) Fuel gauge

Figure 2. Continuous confidence displays.

Interval displays

In statistics, one often measures confidence through the width of a confidence interval. The narrower the interval is, the more certain the system is about the result. We adopt this intuition to recommendation displays. These displays are more informative than both the discrete and the continuous displays presented above as they present information about a predicted score (e.g., recommendation likelihood) and confidence over the same figure.

(a) Simple interval (b) Stars interval

Figure 3. Interval displays.

Simple interval: Figure 3(a) presents the interval in which the predicted score may lie directly. For example, in this figure the score can be anywhere between 1 and 2.5. When the system is confident, the interval becomes a point, displayed by a vertical line. This display is commonly used when reporting confidence intervals in graphs.

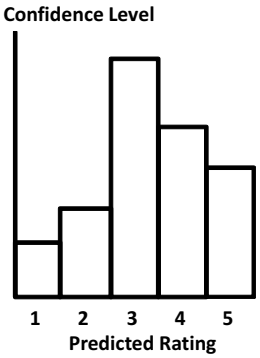
Stars interval: The intuition behind Figure 3(b) is that the predicted score can be anywhere within the colored area. For example, in this figure the score can be anywhere between 2.5 and 4.5. When the system is confident, only a single predicted star should be colored. This display is a variant of the well-known 5-star display for ratings, and is thus reasonably well-known and can be associated with movie recommendations.

Rich displays

Our final set of displays present additional information concerning the confidence level of the prediction. These displays convey information about the likelihood of all possible predictions, that is, how confident is the system in all possible predictions. Such displays may be preferred in cases where a user would like to get as much information as possible through a compact graphical interface.

Predicted Rating	Confidence Level
★	10%
★★	25%
★★★	65%
★★★★	70%
★★★★★	50%

(a) Confidence table



(b) Confidence bars

Figure 4. Rich confidence displays.

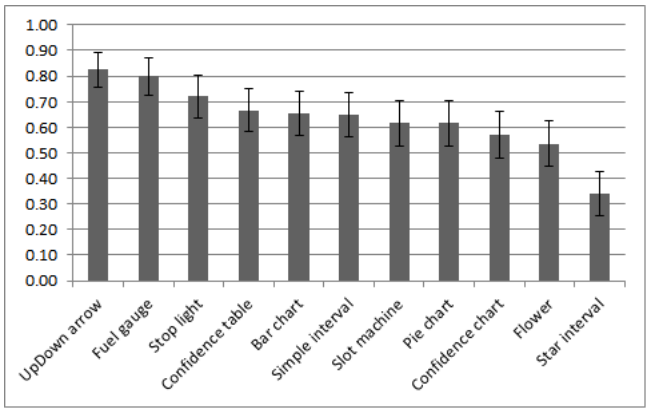
Stars table: Figure 4(a) shows for every possible rating, how confident the system is in that prediction. In the example figure, the system will make a prediction that a rating of 1 star is appropriate with 0.1 confidence. A rating of 5 stars can be given in this example with 0.7 confidence. This display uses the familiar 5 star scale with an added numeric confidence score.

Confidence bars: Figure 4(b) presents the same information but uses a more graphical representation for confidence. While in the stars table ratings were represented graphically using stars, and numbers represented confidence, in the confidence bars display the confidence is represented graphically using bars (where higher means more confident), while the ratings are represented using numbers. The intuition is that, given that some data is represented graphically and some numerically, numeric ratings may be easier to grasp by common users when compared to numeric confidence scores.

User Study

We conduct a user study to estimate the display power, i.e., the understandability and simplicity of each of the suggested displays above. Subjects were shown a series of queries. Each query consisted of a recommendation of a song, given another song, and a confidence display. Subjects were asked two questions: 1)"is the system confident in its recommendation?", with possible answers "yes", "somewhat", "no", and

Figure 5. Success rates for different displays. Error bars represent the adjusted Wald interval at the 0.95 significance level.



2) "how confident is the system in its recommendations?". Answers to these questions depended on the displayed confidence level. For example, if the confidence level was 0.7, the bar chart displayed 4 full bars. Then, possible answers could be, "less than 60%", "between 60% and 80%", and "more than 80%".

The study was designed such that the system first picks a confidence (e.g. 0.8), constructs an appropriate display for it (e.g., 4 full bars in the bar display), and then asks a relevant question (e.g. "is the system more than 70% certain that this is a good recommendation?"). Thus, in the study we always know the correct answers, and we can compare the user's response to the true answer to know whether the user was correct. To avoid confusion, we never choose borderline questions (such as displaying a 0.8 confidence and asking whether the system is more than 80% certain).

After answering queries over all displays, subjects were given an explanation of the displays and were asked to perform the evaluation again (with different items and confidence levels), allowing us to see whether certain displays require training to understand. We had 32 subjects participating in the study, with 26 of them completing both parts (before and after the explanations). The subjects were 4th year engineering student volunteers with a mixed population of males and females and similar ethnic background (all being Israelis of a similar age). We did not collect further demographic details over the test subject in order to reduce the effort and the privacy concerns in participating in this experiment.

All subjects were shown all possible displays in a random order, in a within-subject setting.

We begin by showing the success rates (number of correct answers divided by the number of questions) over the questions for all display types in Figure 5. The up-down arrow

and the fuel gauge were the displays with the highest success rates. That is, most subjects understood the confidence which was presented using these displays and correctly answered the questions. Differences between these two displays and the rest are statistically significant (< 0.011 using a paired T-Test[4]). The display that was most difficult to understand was the star interval (< 0.005 using a paired T-test). We speculate that this is because this display may confuse users due to its similarity to the regular 5-stars rating display.

A deeper look shows additional differences between the displays. Figure 6 provides the complete data over success rates (denoted SR) before and after the explanations (denoted Before and After), for both questions (denoted Q1 and Q2) and the aggregated success rates. There were only 3 displays where the improvement following the explanations was statistically significant; fuel gauge, confidence table, and star interval ($0.013, 0.005, 0.033$ respectively, using a paired T-test). The two latter displays were among the most complicated displays, but the results for fuel gauge are rather surprising. We note, though, that for the fuel gauge, users had difficulty answering the first question ("is the system confident") but not the second question ("how confident is the system"). We speculate that this is because users have different views on translating from a continuous range to a discrete one. We see the reverse effect, difficulty in translating discrete to continuous, in the stop light display before the explanations were presented.

While users could have entered some free text comments, we did not collect any interesting and worthwhile comments.

To conclude, as expected, simpler representations (e.g. the up/down arrow) are easiest to understand and require no explanations. Displays based on well known real world interfaces, such as the fuel gauge, the bar chart, and the stop light, are also easy to understand and require little or no explanation. As we require a reasonable level of expressive power to display sufficient levels of confidence (at least for the purpose of the second study), the bar chart and the fuel gauge seem to be the most appropriate of all the candidates.

Confidence Display Effect on User Behavior

Next we investigate the influence of the confidence displays on user behavior. Confidence displays are designed to provide users with additional data to assist them in making decisions on whether the items recommended to them are truly relevant. We hence check whether users find it easier to identify the quality of the recommended items with or without confidence displays.

To accomplish this, we ran a user study asking users to judge the relevance of several recommendations. The recommendations were either accompanied with a confidence display or not. Below we detail the setup of the experiment and later discuss the results.

Experiment Setup

The user study was implemented as a web-based application where users went through a series of screens, each showing a single query (i.e., whether a set of recommendations is appropriate for an item). The link to the study was sent to students in various phases of their studies, ranging from third year of a B.Sc. to PhD students (the majority being B.Sc. students). We did not collect demographic details over the participants so as to reduce the discomfort and privacy concerns of test subjects. Given this online anonymous setup we could not perform interviews with subjects in order to hear about their subjective opinion of the confidence displays[5], and we leave this to future research.

We asked users to judge the quality of several recommendations in two domains; recommendations of tags for images and recommendations of related movies. In both domains the users were presented with 3 recommendations and were asked to rate each recommendation as either "relevant", "somewhat relevant", or "not relevant". The recommendations were of different quality; at least one recommendation was very relevant and at least one recommendation was not relevant. The third recommendation was sometimes relevant and sometimes of marginal relevance. In some cases the recommendations were accompanied by a confidence display and in some cases not.

On the first page, users received a short explanation asking them to rate the quality of the recommendations and, as a side note, a statement that in some cases the system's confidence in its prediction will be presented. Users were thus unaware of the true goal of the user study. We used a within-subject user study in order to allow us to compare different alternatives for the same user. That is, each user was presented with multiple confidence displays in both domains.

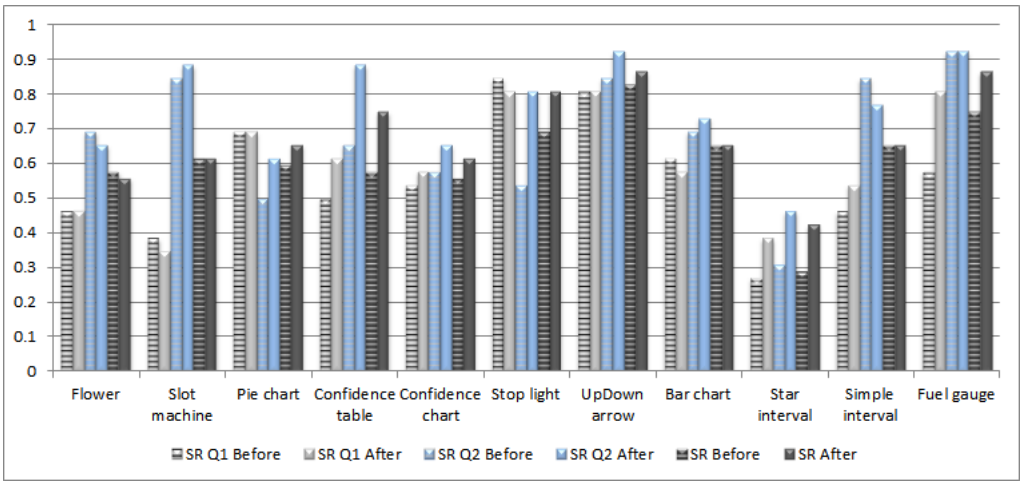
The users were asked to rate at least 15 recommendations in order to be eligible to participate in a raffle but could rate more if they wanted. At the end of the experiment session, users were asked to answer a few questions concerning the displays. Throughout the experiment we alternated the order of the questions, recommendation qualities, and confidence displays randomly between different users. Figure 7 shows an example of an image tag recommendation query.

More formally, the independent variables in this test were the types of display, the confidence level, and the quality of recommendations, while the dependent variable is the perceived recommendation quality. As each user was shown a random ordering of the various domains and types, this

[4]The T-test, here, and throughout this section was executed over the success rate, i.e., the number of correct answers of all displayed questions for each pair of confidence displays.

[5]As with the first study, we allowed users to enter free text comments at the end of the study, but this did not result in any interesting data.

Figure 6. Success rates, before and after explanations



is clearly a within-subjects study. That being said, as we evaluated different ordering of the displays, about half of the users first saw no confidence display while the rest saw confidence displays immediately in the first query. We later discuss some differences between the behavior of the two groups when it comes to trusting the display.

We hypothesis that the type and score of the confidence display influences the perceived quality, in that they help users to better acknowledge the true quality of a recommended item.

Domains. As we explained above, we presented participants with 3 different recommendations of varying quality. We now explain how these recommendations were generated for the two domains. Keep in mind that we are not truly evaluating different recommendation algorithms, only items with different degrees of relevance.

Image tagging: Many professional and non-professional photographers share their photos in public datasets for users to download and use. Retrieval of images is typically based on tags annotating the image. It is in the best interest of the photographer to tag her images as best as possible to help users find them. In such cases, a tag recommendation system, that given an image suggests relevant tags can be highly useful (Sigurbjörnsson & Zwol, 2008).

For this application we used the NUS-WIDE (National University of Singapore Web Image Database) dataset[6] (Chua et al., 2009), containing tags for images. We asked the users to rate the quality of 3 different lists of tags for a displayed image. The lists were generated using the following algorithm: we generated a high quality list by simply picking a random subset of the given tags for the image. As the images in the NUS-WIDE dataset were tagged by humans, these lists of tags are assumed to be of high quality.

We generated a random list by choosing a random subset of the image tags, and then replaced each tag with another tag that frequently co-occurs with it in images, but is not a part of the displayed image tags. We thus simulate a non-relevant, yet not completely random recommendation. We also generated a third list by a combination of the two methods, i.e., choosing 2 of the 3 tags and replacing them with frequently co-occurring tags, while maintaining the third, relevant tag.

Movies: A second well-known problem is the recommendation of movies to watch. For example, a user inserting a movie to her queue in the online video rental and streaming service NetFlix[7] is shown other movies that might be interesting for her, given that single movie. This is the well known item-item scenario which is very popular in the industry (Linden et al., 2003). This is the second domain that we experiment in.

In this domain, we used the MovieLens dataset[8], which contains user ratings for movies. We used a simple item-item algorithm (Linden et al., 2003) to compute for each movie a set of 3 relevant movies. We generated a relevant recommendation by picking a movie with a high item-item score and a non-relevant movie by selecting a movie from the same genre with a low item-item relevance score. The less relevant movies are picked from the same genre in order to maintain some relevance, as a reasonable recommender system is unlikely to present completely irrelevant items. In addition, we showed other recommendations, sometimes relevant and sometimes non-relevant.

In the movie domain, users also had the option to click on a movie which opened the related Internet Movie Database (IMDB) page[9], in case they needed information about the movie in order to make a decision. There were 217 information requests on movies out of 4,548 movie displays (each movie could be presented several times for different users).

[6] lms.comp.nus.edu.sg/research/NUS-WIDE.htm
[7] www.netflix.com
[8] www.grouplens.org
[9] www.imdb.com

Figure 7. Judging the quality of different recommended tags lists for an image.

There weren't more information requests partially because we limited the domain to the 200 most popular movies in order to reduce the chance that participants did not recognize the movie.

We also added an option "I don't know this movie" for either the queried movie or the recommended movies, as well as a "I don't know" option for image tag lists. There were 696 selections of "don't know" options.

In general, relevance is subjective, e.g., it is difficult for a user to estimate whether a specific item is relevant to her peers. On the other hand, in the tasks that we explore, relevance is with respect to a specific item. In the image domain, the tag "beach" is clearly appropriate for an image showing the ocean shore. There can still be some subtle differences between users, but in general, relevance is quite objective in this domain. This is true to some extent in the movie domain too. The user preference over the movie does not influence her opinions concerning which movies are relevant to it. That being said, a variety of subjective opinions concerning movie-to-movie relevance is indeed possible.

Confidence Computation and Display. To compute confidence for the recommendations, we randomly selected a confidence score for different intervals; a high confidence interval and a low confidence interval. For example, for a 6 category confidence display, the high confidence was either 5 or 6, and the low confidence was either 2 or 3. We never reported a lower confidence because it does not make sense for a system to recommend items of which it is not certain at all. For the relevant recommendation we randomly selected a confidence from the high confidence interval and for the non-relevant recommendation, we randomly selected a confidence from the low confidence interval. Using this method, we ensure that the difference in confidence between items in the same lists is bounded, making it easier to judge the quality of the entire list.

We limited the user study to 3 confidence displays; Bar chart, Fuel Gauge, and Star interval. The Fuel gauge was selected because it is the continuous display with the highest accuracy (see Figure 5). For the discrete presentation, we selected the Bar chart (rather than the more accurate Up-Down arrow) because it is rich enough to display sufficient confidence categories. Finally, we selected the Star interval display because it had the lowest accuracy, and was useful in contrasting user's behavior for good and bad confidence displays.

Results

We had 160 participants (mostly students) who volunteered to participate in the study, 21 of which did not complete the mandatory 15 queries and left without answering the questions at the end of the survey. The participants received an email asking them to participate in a user study and enter a raffle to win a digital camera. The study was done online and was accessed through a link attached to the email. In total, the participants rated 10,945 recommendations, 6,397 image tag recommendations, and 4,548 movie recommendations. On average, each user answered 22.8 queries (related to movies or relevant tags for an image).

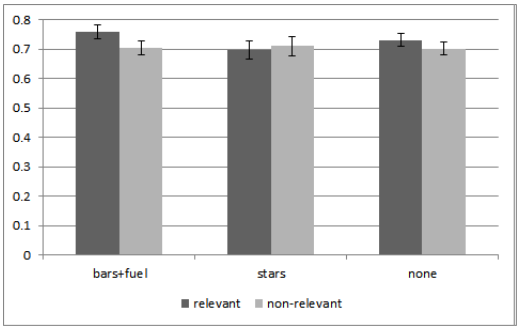
All subjects were asked questions for all display types, in a within-subject setting. The order of the displays was randomized

When computing statistical significance, we treated each user as independent, i.e., aggregated over the observations for each user and running statistical significance over the per-user aggregated observations. For example, if a user answered 10 queries, and was correct in 8 of her judgments, we compute her success rate to be 0.8. In the discussion below wherever we discuss a difference between two methods, it was found to be statistically significant. When we state that differences were minor they were also found not to be statistically significant.

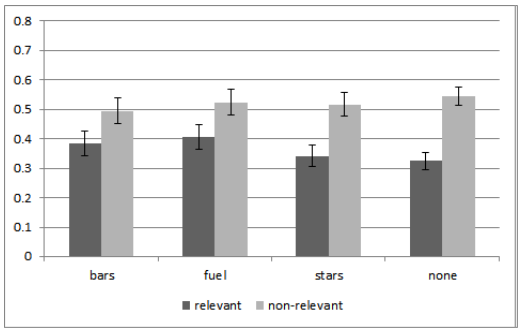
Recommendation Quality Identification. We first examine the hypothesis that users find it easier to identify the quality of the recommendation when confidence is displayed. Figure 9(a) and Figure 9(b) show the accuracy of the identification. Accuracy is measured as the portion of correct answers (i.e., "relevant" for a relevant recommendation and

"not-relevant" for a non-relevant recommendations) from the total number of answers. As we can see, differences between the different confidence displays (or lack there of) are minor and in some cases not statistically insignificant. In the image domain (Figure 9(a)) users easily identified the relevant tags and the non-relevant tags, with or without confidence. The complicated stars display slightly reduced performance even when compared to no display, as expected.

Figure 8. Accuracy of identification of recommendation quality. Error bars represent the adjusted Wald interval at the 0.95 significance level.



(a) Image tag lists



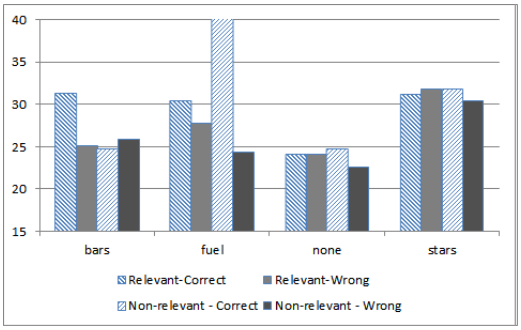
(b) Recommended movies

In the movie domain (Figure 9(b)) users had a lower accuracy in identifying relevant and non-relevant movies. This is partially because in this domain we used an actual recommendation technique. Thus, some movies that the algorithm considered as relevant may not seem so to others. That being said, deciding whether two movies are related is a more subjective question and opinions on it may vary more than for adequate tags for an image. In this domain, identifying non-relevant movies was much easier for users. Still, in this domain as well, the differences between confidence displays are minor and not statistically significant.

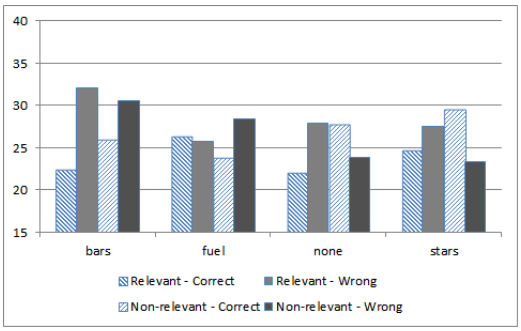
We also checked whether displaying a confidence affected the time that users needed in order to decide whether recommendations were relevant. Figure 10(a) and Figure 10(b) show the average time that users needed before making their choices for both domains. In the image domain, users in general needed less time when no confidence display was presented. This is most notable in the case of correctly identifying relevant recommendations (more than 30 seconds for all displays and less than 25 seconds given no display), which may be considered the easiest task. This may be attributed to the additional cognitive load when the user is presented with two sources of information, the recommendation and the display, and must process then both. The user must spend an additional mental effort to resolve the conflict, especially in the case when the displayed confidence is different than the users' perceived quality. This behavior was repeated, to some extent, in the movie domain. These figures also do not provide any solid evidence that confidence displays helped users in identifying the recommendation quality, although they do show that people notice and dwell on the confidence displays in many cases.

A similar increase in response time when presenting confidence scores was also observed by Rukzio, Hamard, Noda, and De Luca (2006) although a different application with a different confidence visualization technique.

Figure 9. Average response time. Standard deviation was less than 0.001 in all cases and was thus omitted from the graphs.



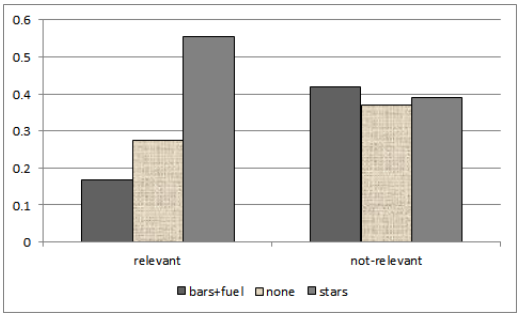
(a) Image tag lists



(b) Recommended movies

Finally, we took a closer look at cases where users said that they "don't know" the correct answer in Figure 10. We can see that for relevant recommendations, users were less confused about the quality of the recommendation when shown the bars or the fuel display, or even when no confidence score is presented, than when shown the stars display.

Figure 10. Portion of "don't know" answers for different recommendation quality and different displays in the image tags domain. The stars display was significantly worse then the alternatives for relevant items with a p-value less than 0.05 using a χ^2 test.



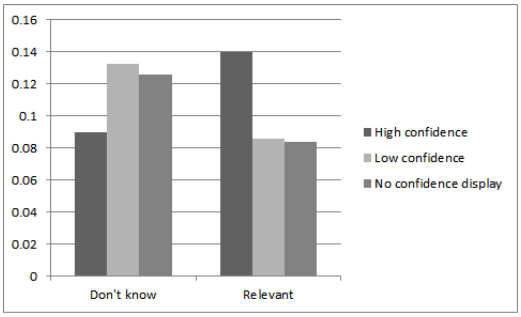
This indicates that the confusing, over complicated stars display caused people to be less sure of their opinion. On the other hand, the bars and the fuel displays caused people to be more certain and less likely to say that they "don't know" the correct answer when the recommendation is relevant. When recommendations were not-relevant, a different phenomenon exists; users are equally unsure given any confidence display or lack thereof. That is, when people see a non-relevant recommendation of whose quality they are unsure, they do not tend to believe the confidence display, yet when they see a relevant recommendation that they are unsure of, they are more likely to trust the system confidence display, provided that it is easy to understand.

These results were not reproduced in the movie domain where no statistically significant difference was found. We speculate that this is because for movies, people tend to mark "don't know" when they are unfamiliar with the movie and hence a confidence display cannot help nudging them towards making a decision. In the image tags domain people may be able to judge whether a certain tag is adequate for an image relatively easily. This may be because image tagging is a familiar task for users, e.g., in Flickr[10] or Facebook[11], or because reasonable tag relevance decisions can be made based purely on common sense.

Believing the Confidence Displays. We also checked whether users trust the confidence displays, e.g., rate a non-relevant recommendation as "relevant", when the system reports high confidence in the recommendation. This is similar to questions that were examined by Cosley et al. (Cosley, Lam, Albert, Konstan, & Riedl, 2003), who concluded that users can be manipulated into giving other ratings than they originally intended given lower or higher displayed predicted ratings.

To measure this, in some cases (10% of the queries), a weak recommendation received a high confidence display. As Figure 11 illustrates, users indeed rated such recommendations more often as "relevant". Furthermore, the number of times that users answered "don't know" is reduced by almost the same amount. A possible explanation is that when users are unsure, and a high confidence is presented to them, they tend to believe the system. We believe that the bias of people's judgment of the recommendations given high confidence displays can be attributed to the well-studied "anchoring" phenomena (Kahneman, Slovic, & Tversky, 1982), where people's estimations are influenced by an anchor which was presented to them prior to the estimation. The anchoring effect was also studied with similar observations when artificial rating predictions are displayed to users (Adomavicius, Bockstedt, Curley, & Zhang, 2011). In confidence displays, a similar phenomenon was also observed by Antifakos, Kern, Schiele, and Schwaninger (2005) in a different application, where the user trust in the system increased when confidence scores were presented.

Figure 11. Portion of user ratings for non-relevant recommendations, given different confidence displays. Differences between high confidence and the other two alternatives are significant with a p-value of less than 0.05 using χ^2 test.



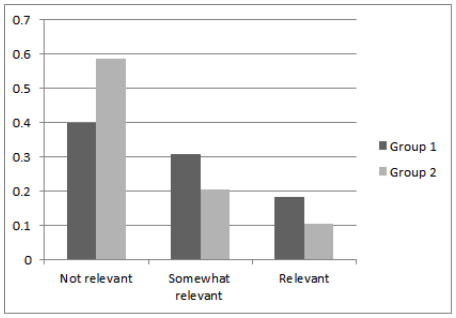
In the experiment, users were also uniformly split into two groups. Group 1 received confidence displays immediately, while group 2 started viewing confidence displays only halfway through the experiment. Thus, we can say that group 2 was "trained" to evaluate the recommendations directly before observing any confidence display, while group 1 was "trained" to consider the confidence displays, although there was no formal training phase. As Figure 12 shows, there is a considerable difference between the two groups when we present a high confidence for a non-relevant recommendation. Users from group 2 evaluated the recommendation on face value, and concluded that it is not relevant, while users from group 1 tended to believe the confidence display more [12].

[10] urlwww.flickr.com

[11] urlwww.facebook.com

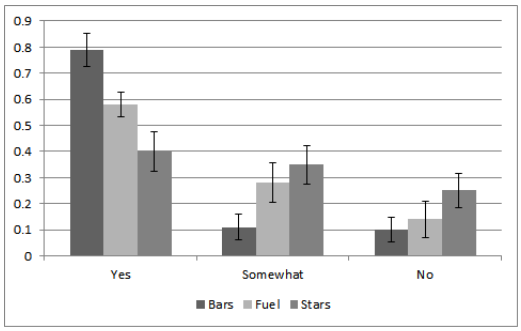
[12] In the previously reported results both groups were evaluated together, because there was no statistically significant difference between the groups. For example, we can't say that group 1 achieved a higher accuracy when judging relevance of movies.

Figure 12. Portion of user ratings for non-relevant recommendations with a high confidence display. Differences between the groups are significant with a p-value of less than 0.04 using χ^2 test.



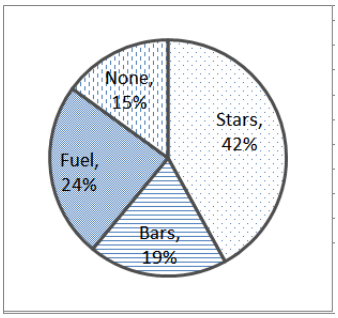
Participants Perception. At the end of the trial, participants were asked whether the confidence display helped them in accomplishing the task. 42% of the participants answered that the confidence display did not help, while 22% answered that it did help. 36% answered that the display "somewhat" helped them. Given the participantâĂŹs performance with and without confidence displays, it is surprising that many participants perceived some value in the displays at all.

Figure 13. Did the participants understand the display? Y axis is the participants portion. Error bars represent the adjusted Wald interval at the 0.95 significance level.



We also asked participants whether they understood the different confidence displays that were presented to them. Figure 13 shows the results for these questions. As expected, the stars display was the least understandable for participants. During the trial, participants could also ask for an explanation page over the current query. The explanation page contained explanations over all elements, including the query, the recommendations, and the confidence displays. Only 37 participants asked for explanations and only 10 of those asked for more than a single explanation, suggesting that the confidence displays were reasonably intuitive. The understandability of the various displays can also be measured by looking at the displayed confidence when participants ask for explanations. As Figure 14 shows, 42% of the requests came when the stars display was used, providing further evidence that is was less understandable than the rest of the displays, as expected.

Figure 14. Requests for explanations.



Discussion

In contrast to our hypothesis, the confidence displays did not ease the task of identifying the recommendation quality (except for the reduction in "don't know" answers). In part, this can be attributed to the limitations of the user study; in the domains we used, people could, in principle, form an opinion concerning the quality of the recommendation fairly easily in most cases (less so for the movie domain). It is not clear whether our results would carry into domains where people must select and use the item before they know whether it is appropriate. For example, in recommending hotels, it is possible that people will prefer a hotel that the system is confident they will like because they can only know whether the hotel was truly appropriate choice after visiting it. The same may also hold for reading news stories. It seems, though, that people tend to trust the system confidence especially when they are in doubt. In such domains, hence, it might be that confidence displays can truly help people in making decisions. This conclusion is supported by Cosley et al. (Cosley et al., 2003) who also find that the display of information about a recommendation influences the way that people act.

Users who interact with a top-N recommender system typically receive personalized recommendations, i.e., recommendations based on items that they already prefer. For example, in a news website such as CNN.com, the recommendations for "related stories" appear below the current story. Typically users view the recommendations after having read a story, indicating a certain interest in the content. In this user study, however, users were presented with recommendations given arbitrary items that they may like, dislike, or may be indifferent about. We speculate that the results may not be drastically different in an online study because the confidence displays concern only the recommended items. These recommendation may always be interesting or not interesting for the active user, even in a personalized setting. We leave a

true online experiment with real users of a real system using items they prefer that can verify our speculations to future research.

While the goal of the experiment was to evaluate the effects of confidence displays, the participants were not informed of our agenda. In informal talks with some test subjects it became apparent that they did not understand that the focus of the study was upon the displays and did not realize that we were not actually comparing the quality of different recommendation algorithms.

A different yet related question is which displays are most convenient for users in providing ratings over items to the system. Sparling and Sen (Sparling & Sen, 2011) suggested four different displays with various level of granularity, from unary "like" votes, through binary and five-star ratings, to 100 scale slider. They allowed users to express their opinion over movies and product reviews. Their main conclusions are that users require more time when using a finer-grained ratings display. They also show that discrete, well-known presentations (the five-stars and the thumbs up-down binary presentation) being preferred by users. Gena *et al.* (Gena, Brogi, Cena, & Vernero, 2011) use s similar setting with a set of rating scales, and show that different scales induce different ratings, that is, that user opinion is influenced by the scale that is used.

We suggest presenting two different results to the user — the list of recommended items, and the confidence of the system over the items in the list or the list itself. While providing feedback over the recommended item (e.g. "like", or "not interesting") is natural, providing feedback over the confidence results does not seem meaningful. Still, it seems reasonable that a system deciding on a display should be consistent in the two tasks — presenting the system confidence and allowing users to express feedback over the recommended items. Thus, it may be appropriate to chose a single display taking into a account both the consideration discussed in this paper as well as the considerations discussed by Sparling and Sen.

Limitations

There are some limitations in the user study we conducted that may limit our ability to draw conclusions from the results.

First, the study was not done within the scope of a real system with real users attempting to fulfill their goals. It is possible that user behavior with respect to the confidence displays might be different when users browse items of true interest and utility. We are currently in the process of evaluating confidence displays on a real system offering online video games. On the other hand, as the study participants were all volunteers, with little to gain from participation, we made the conscious decision to reduce the number of questions to a minimum. As such, we did not ask users a large number of post-study questions which could have revealed more interesting aspects of user behavior. It is possible that a smaller lab study with paid participants who are obligated to answer more questions could help us to better understand how users respond to confidence displays.

Second, there are a vast number of possible displays that we have not investigated in this work. Clearly, it is possible that other displays could have been more intuitive, useful, and helpful to users. We hope that this work will encourage others to develop and evaluate more confidence displays. Symbols are also interpreted differently by people from different cultures. As the ethnic and cultural background of our users was similar, we cannot draw conclusions about people from different backgrounds.

Conclusion

Some recommender systems can compute their confidence in the recommended items in addition to a list of recommendations. In this paper we studied questions associated with the display of the system's confidence in its recommendations.

We presented a set of possible confidence displays which varied in their performance, their complexity, the type of information they presented, and the knowledge required to understand the display. We ran a user study to compare the various displays, showing that some displays are more understandable and are better liked by users. Most notably, users best understood the displays that were inspired by well-known displays in other areas, such as the Bar Chart presentation often associated with cell phone connectivity, and the Fuel Gauge inspired by fuel gauges in cars.

In a second user study we investigated the effect that a confidence display has on users when they evaluate recommended items. While it might be assumed, à priory, that such a display can help users identify relevant items, our study does not provide much evidence to support this claim. Users require more time in general when confidence displays are shown, and gain no significant accuracy improvement over instances where no confidence is displayed. The study evaluated two different domains; movies and image tags. It may be that in other domains, where the quality of the presented items cannot be directly assessed without experiencing it[13], confidence displays may provide more value to the user.

The main findings of our experiments are:

- Discrete confidence displays with a relatively low number of scores are most understandable to users, requiring no training. Displays based on well-known interfaces are easier to understand.
- Confidence displays did not help users in identifying relevant items in general. There was no improvement in the

[13]This is also true in the case of the movie domain where people don't know the recommended movie. In our study, however, subjects were asked to skip recommendations for movies they were unfamiliar with.

accuracy of the identification, nor in the required time for making a decision. Users turn to the confidence display when they are unsure whether a relevant item is truly relevant.

• Overall confidence displays are natural to users, and people understand them without requiring a training period.

• Users build trust over the confidence displays, and thus confidence displays can steer users towards less relevant items by displaying high confidence in the item's relevance to the user.

• Trust is built over time — users require a few interactions with the system before they are willing to believe its confidence.

The last two finding should be taken with a grain of salt, as these were only the conclusion of a short user study and not thorough interactions with a real recommendation system. It is likely that with real systems people will grow to dislike the system if it is often presenting miscalculated confidence in its predictions.

In the future we intend to study different cases in order to see whether confidence displays can help users in other domains. Specifically, it would be interesting to examine news stories where, although the title provides some information over the item's relevance, users must read the story before forming a concrete opinion. It might be that in such domains, people will more often select items with a higher reported confidence.

References

Adomavicius, G., Bockstedt, J., Curley, S., & Zhang, J. (2011). Recommender systems, consumer preferences, and anchoring effects. In *Recsys'11 workshop on human decision making in recommender systems.*

Antifakos, S., Kern, N., Schiele, B., & Schwaninger, A. (2005). Towards improving trust in context-aware systems by displaying system confidence. In *Proceedings of the 7th international conference on human computer interaction with mobile devices & services* (pp. 9–14). New York, NY, USA: ACM.

Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the acm international conference on image and video retrieval* (pp. 48:1–48:9). New York, NY, USA: ACM.

Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., & Riedl, J. (2003). Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 585–592). New York, NY, USA: ACM.

Delany, S. J., Cunningham, P., Doyle, D., & Zamolotskikh, A. (2005). Generating estimates of classification confidence for a case-based spam filter. In *Proceedings of the 6th international conference on case-based reasoning research and development* (pp. 177–190). Berlin, Heidelberg: Springer-Verlag.

Desrosiers, C., & Karypis, G. (2011). A comprehensive survey of neighborhood-based recommendation methods. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook* (p. 107-144). New York, NY, USA: Springer.

Dredze, M., Crammer, K., & Pereira, F. (2008). Confidence-weighted linear classification. In *Proceedings of the 25th international conference on machine learning* (pp. 264–271). New York, NY, USA: ACM.

Esposito, F., Malerba, D., & Semeraro, G. (1997, May). A comparative analysis of methods for pruning decision trees. *IEEE Trans. Pattern Anal. Mach. Intell.*, *19*(5), 476–491.

Gena, C., Brogi, R., Cena, F., & Vernero, F. (2011). The impact of rating scales on user's rating behavior. In *Proceedings of the 19th international conference on user modeling, adaption, and personalization* (pp. 123–134). Berlin Heidelberg: Springer.

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 acm conference on computer supported cooperative work* (pp. 241–250). New York, NY, USA: ACM.

Herman, L., Ornstein, G., & Bahrick, H. (1964). Operator decision performance using probabilistic displays of object location. *Human Factors in Electronics, IEEE Transactions on*, *HFE-5*(1), 13-19.

Hofmann, T. (2003). Collaborative filtering via gaussian probabilistic latent semantic analysis. In *Proceedings of the 26th annual international acm sigir conference on research and development in informaion retrieval* (pp. 259–266). New York, NY, USA: ACM.

Kadie, C. M., Meek, C., & Heckerman, D. (2002). Cfw: A collaborative filtering system using posteriors over weights of evidence. In *Uai '02, proceedings of the 18th conference in uncertainty in artificial intelligence* (p. 242-250). Burlington, MA, USA: Morgan Kaufmann.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. New York, NY: Cambridge University Press.

Karatzoglou, A., & Weimer, M. (2010). Quantile matrix factorization for collaborative filtering. In *E-commerce and web technologies, 11th international conference, ec-web 2010, bilbao, spain, september 1-3, 2010. proceedings* (Vol. 61, p. 253-264). New York, NY, USA: Springer.

Katz, G., Shabtai, A., Rokach, L., & Ofek, N. (2012). Confdtree: Improving decision trees using confidence intervals. In *12th ieee international conference on data mining, icdm 2012, brussels, belgium, december 10-*

13, 2012 (p. 339-348). Washington, DC, USA: IEEE Computer Society.

Konstan, J. (2010). *HCI for Recommender Systems: An Introduction.* Tutorial at the ACM Recommender Systems.

Krzanowski, W. J., Bailey, T. C., Partridge, D., Fieldsend, J. E., Everson, R. M., & Schetinin, V. (2006). Confidence in Classification: A Bayesian Approach. *Journal of Classification*, *23*, 199–220.

Linden, G., Smith, B., & York, J. (2003, January). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76–80.

Littman, M. L., Keim, G. A., & Shazeer, N. (2002, January). A probabilistic approach to solving crossword puzzles. *Artif. Intell.*, *134*(1-2), 23–55.

McNee, S. M., Lam, S. K., Guetzlaff, C., Konstan, J. A., & Riedl, J. (2003). Confidence displays and training in recommender systems. In *International conference on human-computer interaction INTERACT'03* (p. 69-74). Amsterdam, the Netherlands: IOS Press.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 152–158). New York, NY, USA: ACM.

Rokach, L., Naamani, L., & Shmilovici, A. (2008, October). Pessimistic cost-sensitive active learning of decision trees for profit maximizing targeting campaigns. *Data Mining and Knowledge Discovery*, *17*(2), 283–316.

Rukzio, E., Hamard, J., Noda, C., & De Luca, A. (2006). Visualization of uncertainty in context aware mobile applications. In *Proceedings of the 8th conference on human-computer interaction with mobile devices and services* (pp. 247–250). New York, NY, USA: ACM.

Sigurbjörnsson, B., & Zwol, R. van. (2008). Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on world wide web* (pp. 327–336). New York, NY, USA: ACM.

Sparling, E. I., & Sen, S. (2011). Rating: how difficult is it? In *Proceedings of the fifth acm conference on recommender systems* (pp. 149–156). New York, NY, USA: ACM.

Tintarev, N., & Masthoff, J. (2011). Designing and evaluating explanations for recommender systems. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook* (p. 479-510). New York, NY, USA: Springer.

Toth, N., & Pataki, B. (2007). On classification confidence and ranking using decision trees. In *11th international conference on intelligent engineering systems. ines'07.* (p. 133-138). Piscataway, NJ, USA: IEEE Conference Publications.