How should I explain? A comparison of different explanation types for recommender systems

Fatih Gedikli[a], Dietmar Jannach[a], Mouzhi Ge[b]

[a]*TU Dortmund, 44221 Dortmund, Germany*
[b]*Bundeswehr University Munich, 85579 Neubiberg, Germany*

Abstract

Recommender systems help users locate possible items of interest more quickly by filtering and ranking them in a personalized way. Some of these systems provide the end user not only with such a personalized item list but also with an explanation which describes why a specific item is recommended and why the system supposes that the user will like it. Besides helping the user understand the output and rationale of the system, the provision of such explanations can also improve the general acceptance, perceived quality, or effectiveness of the system.

In recent years, the question of how to automatically generate and present system-side explanations has attracted increased interest in research. Today some basic explanation facilities are already incorporated in e-commerce Web sites such as Amazon.com. In this work, we continue this line of recent research and address the question of how explanations can be communicated to the user in a more effective way.

In particular, we present the results of a user study in which users of a recommender system were provided with different types of explanation. We experimented with ten different explanation types and measured their effects in different dimensions. The explanation types used in the study include both known visualizations from the literature as well as two novel interfaces based on tag clouds. Our study reveals that the content-based tag cloud explanations are particularly helpful to increase the user-perceived level of transparency and to increase user satisfaction even though they demand higher cognitive effort from the user. Based on these insights and observations, we derive a set of possible guidelines for designing or selecting suitable explanations for recommender systems.

Keywords: Recommender Systems, Decision Support, Explanations, Interface Design, Collaborative Filtering, Tag Clouds, User Evaluation

1. Introduction

Recommender systems point online users to possibly interesting or unexpected items, thereby increasing sales or customer satisfaction on modern e-commerce platforms [Linden et al., 2003; Senecal and Nantel, 2004; Zanker et al., 2006; Dias et al., 2008; Jannach and Hegelich, 2009].

Email addresses: fatih.gedikli@tu-dortmund.de (Fatih Gedikli), dietmar.jannach@tu-dortmund.de (Dietmar Jannach), mouzhi.ge@ebusiness-unibw.org (Mouzhi Ge)

However, personalized recommendation lists alone might be of limited value for the end users when they have to decide between different alternatives or when they should assess the quality of the generated recommendations. In other words, only showing the recommendation lists can make it hard for the users to decide whether they can actually trust that the recommended items are actually useful and interesting without inspecting all of them in detail.

One possible approach to support the end user in the decision making process and to increase the trust in the system is to provide an explanation for why a specific item has been recommended [Herlocker et al., 2000; Bilgic and Mooney, 2005; Pu and Chen, 2006; Tintarev and Masthoff, 2007a,b; Friedrich and Zanker, 2011]. In general, there are many approaches of explaining recommendations, including non-personalized as well as personalized ones. An example of a non-personalized explanation would be Amazon.com's *"Customers who bought this item also bought..."* label for a recommendation list, which also carries explanatory information.

This work deals with questions of how explanations could be communicated to the user in a more effective way. This includes both questions of the *visual representation* as well as questions of the *content* to be displayed. In general, the type and depth of explanations a recommender system can actually provide depend on the types of knowledge and/or algorithms that are used to generate the recommendation lists. In knowledge-based recommendation or advisory approaches, explanations can be based on the rule base which encodes an expert's domain knowledge and the explicitly acquired user preferences [Felfernig et al., 2007; Jannach et al., 2009; Zanker, 2012]. For the most prominent type of recommender systems, collaborative filtering recommenders, Herlocker et al. [2000] and Bilgic and Mooney [2005] have proposed various ways of explaining recommendations to the user. Herlocker et al. have also shown that explanations can help to improve the overall acceptance of a recommender system.

In this paper, we continue the line of work of Herlocker et al. [2000], Bilgic and Mooney [2005], Vig et al. [2009], Tintarev and Masthoff [2007a, 2012], and our own work presented in [Gedikli et al., 2011]. In particular, we aim to contribute to the following research questions.

1. The main question we seek to answer in this paper is which effects different explanation types for recommendations have on users. In the existing literature on recommender system explanations, authors often limit their analysis to some specific explanation goals [Herlocker et al., 2000; Bilgic and Mooney, 2005; Pu and Chen, 2006, 2007; Symeonidis et al., 2009] or explanation types [Vig et al., 2009; Gedikli et al., 2011]. In our work, however, we aim at evaluating different explanation types in a comprehensive manner and consider the desired effects and quality dimensions *efficiency*, *effectiveness*, *persuasiveness*, *perceived transparency*, and *satisfaction* [Tintarev and Masthoff, 2011] in parallel. To that purpose, we conducted a laboratory study involving 105 subjects in which we compare several existing explanation types from the literature ([Herlocker et al., 2000]) with a tag-based explanation approach.

2. Going beyond existing research which focuses only on one explanation goal or analyze trade-offs between two quality dimensions[1], we aim at detecting interdependencies between more than two quality dimensions. In particular, our goal is to analyze the influence of efficiency, effectiveness, and perceived transparency on user satisfaction. Based on the dependencies between the different effects of explanation types, we aim to derive a first set

[1] See Table 2 in [Tintarev and Masthoff, 2012].

of possible guidelines for the design of effective and transparent explanations for recommender systems and sketch potential implications of choosing one over the other. These guidelines were validated through a qualitative interview-based study involving 20 participants.

3. We finally aim to obtain a deeper understanding of the value of the recently proposed tag- and preference-based explanation types proposed in [Gedikli et al., 2011]. We included two variants of this explanation method in our experimental study and compare their performance with the other explanation types in the different quality dimensions. Since acquiring explicit tag preferences is costly and can be cumbersome for the user, one of the two tag-based explanations incorporates a new method to automatically estimate the user's detailed preferences from the item's overall ratings.

The paper is organized as follows. Section 2 summarizes the quality factors for recommender system explanations and discusses related and previous work. Section 3 introduces the different explanation types compared in our study. Section 4 describes the experimental setup. Section 5 provides a discussion of the obtained results and our first set of possible design guidelines. Section 6 finally summarizes the main findings of this work and gives an outlook on future work.

2. Explanations in recommender systems

In recent years, the concept of explanations has been widely discussed in the area of recommender systems [Pu and Chen, 2007; Tintarev and Masthoff, 2008; Vig et al., 2009; Friedrich and Zanker, 2011; Tintarev and Masthoff, 2012]. An explanation can be considered as a piece of information that is presented in a communication process to serve different goals, such as exposing the reasoning behind a recommendation [Herlocker et al., 2000] or enabling more advanced communication patterns between a selling agent and a buying agent [Jannach et al., 2010]. Up to now, however, there exists no standard definition of the term "explanation" in the context of recommender systems. According to Tintarev and Masthoff [2012], a popular interpretation of the term explanation in recommender systems is that explanations *"justify"* the recommendations. Since this definition might be too narrow, we propose to characterize explanations through the possible aims which one might want to achieve with them in a recommendation system. Tintarev and Masthoff identify seven possible aims of explanations for recommender systems as shown in Table 1.

(1) Efficiency	Reducing the time used to complete a task
(2) Effectiveness	Helping the users make better decisions
(3) Persuasiveness	Changing the user's buying behavior
(4) Transparency	Explaining why a particular recommendation is made
(5) Satisfaction	Increasing usability and enjoyment
(6) Scrutability	Making the system's user model correctable
(7) Trust	Increasing the user's confidence in the system

Table 1: Possible goals of using explanations in recommender systems.

This paper investigates the impact of different explanation types on the first five factors in this list. Next, we will characterize these factors in more detail and sketch how to measure each of them.

2.1. *Efficiency*

An explanation is usually considered to be efficient when it helps the user to decide more quickly or when it helps to reduce the cognitive effort required in the decision process. In the context of conversational recommender systems, Thompson et al. [2004], for example, measure efficiency by computing the total interaction time between the user and the recommender system until the user has found a suitable item. McSherry [2005], in contrast, measures efficiency through the number of required dialogue steps before a user accepts one of the system's recommendations. In other papers, efficiency is sometimes calculated by measuring the time used to complete the same task with and without an explanation facility or with different types of explanations, see, e.g., the study of Pu and Chen [2006].

In our work, we adopt an efficiency measure that is based on the decision time required by a user. We distinguish between "item-based" and "list-based" efficiency. Typically, recommender systems can produce two types of output: a) a rating prediction showing to what degree the user will like or dislike an item and b) a list of n recommended items. Therefore, efficiency can be measured either for each individual item or for a given list of recommendations. Item-based efficiency thus considers the decision time required by a user to evaluate a single candidate item at a time (see, e.g., [Gedikli et al., 2011]). An appropriate protocol for list-based efficiency would be to measure the overall time required by a user to decide on one single best item given a larger candidate set with explanations (see, e.g., [Thompson et al., 2004; McCarthy et al., 2005]). In order to make the results comparable to our prior work on explanations ([Gedikli et al., 2011]), we decided to measure the item-based efficiency in this work.

2.2. *Effectiveness*

Effectiveness can be defined as the ability of an explanation facility to help users make better decisions. Effective explanations support the users in correctly determining the actual quality or suitability of the recommended items and filter out uninteresting items [Bilgic and Mooney, 2005; Tintarev and Masthoff, 2012]. Overall, such explanations can significantly help to increase the overall utility and usability of a recommender system.

The effectiveness of explanations can be measured in different ways. Vig et al. [2009], for example, performed a user study in which they presented different explanation types to the users and then asked them how well the individual explanation types helped them to decide whether or not they liked an item. Bilgic and Mooney [2005], in contrast, determine effectiveness by measuring the closeness between the user's estimate of the quality or appropriateness of an item and the actual quality or utility of the recommended items.

Similar to Tintarev and Masthoff [2012], we adopt this second approach in our laboratory study and use the following procedure. First, users are asked to estimate the quality of a recommended item by considering only the explanation generated by the recommender. Afterwards, users use or "consume" the item (e.g., watch a movie or analyze the item based on more detailed information) and rate the item again based on their real experiences or the additional knowledge. The closeness of the two ratings can then be used as a proxy to measure effectiveness.

2.3. *Persuasiveness*

Persuasiveness, sometimes referred to as promotion, is strongly related to effectiveness and can be defined as the ability of an explanation type to convince the user to accept or disregard certain items. We can discriminate between overestimate- and underestimate-oriented persuasiveness.

If persuasiveness is overestimate-oriented, the user might overrate the quality or suitability of an item. This form of persuasiveness can be used to manipulate the user's opinion about certain items, e.g., in a sales promotion. Underestimate-oriented persuasiveness is an effect where the explanation lets an item's relevance appear lower than it actually is. Such an effect might be desired when the goal is to direct customers to a certain part of the spectrum, e.g., to niche items. The level and form of intended persuasion should thus be in line with the business strategy.

The level of persuasiveness of an explanation can be approximated by a measurement which determines to which extent the user's evaluation is changed by the explanation [Bilgic and Mooney, 2005]. In our experimental study, we will use such a measure of persuasiveness and consider it together with the effectiveness of an explanation type.

2.4. Transparency

Transparency in a recommender system is related to the capability of a system to expose the reasoning behind a recommendation to its users [Herlocker et al., 2000]. While many recommendation applications today represent a "black box" that accepts inputs and returns recommendation lists, a transparent recommender would also try to explain (parts of) the reasoning behind a recommendation to the user. Historically, such explanations were particularly important in classical knowledge-based systems and helped these system produce more credible predictions or more accountable decision support (see, e.g., [Rowe and Wright, 1993] or [Ong et al., 1997]). In the domain of recommender systems, transparency is considered as an important factor that contributes to users building trust in the system [Swearingen and Sinha, 2002].

We can discriminate between objective transparency and user-perceived transparency. Objective transparency means that the recommender reveals the actual mechanisms of the underlying algorithm. However, there might be a number of reasons why it might be more appropriate to present more user-oriented "justifications" ([Vig et al., 2009]) than to try to explain the rationale of the recommendation algorithm: the algorithm might for example be too complex to explain, not intuitive, or the algorithm details have to be protected. These justifications are therefore often more shallow and user-oriented. User-perceived transparency in that context is thus based on the subjective opinion of the users about how good the system is capable of explaining its recommendation logic[2].

In our laboratory study, we assess the effect of different explanation types based on *user-perceived* transparency. Correspondingly, we rely on questionnaires which the users filled out after interacting with the system[3].

2.5. Satisfaction

The user's overall satisfaction with a recommender system is assumed to be strongly related to the perceived quality of its recommendations and explanations, see [Swearingen and Sinha, 2002], [Cosley et al., 2003], and [McCarthy et al., 2004]. Beside the satisfaction with the system as a whole, one can however also measure the user's quality perception of the explanations themselves. One typical method of measuring this is again to directly ask users to which extent they liked the explanation type (*"How good do you think this explanation is?"*) [Tintarev and Masthoff, 2012].

[2]There are reports in the literature that show that the user-perceived transparency can be high even though the explanation do not actually correspond to the underlying recommendation logic, see [Herlocker et al., 2000].

[3]When using the term transparency in the rest of the paper, we therefore mean user-perceived transparency.

In our experimental study, we are interested in analyzing the users' satisfaction with the explanation types and follow the approach to directly ask the users to rate the explanation types. Furthermore, we will analyze if other factors (efficiency, effectiveness, and transparency) measurably influence the user's satisfaction with the different types of explanation.

3. Overview of the evaluated explanation types

In order to analyze the effects of different explanation types on users, we evaluated ten different approaches. Seven of them were proposed and evaluated in [Herlocker et al., 2000], two are based on tag clouds and have been introduced in our own previous study [Gedikli et al., 2011], and one is based on a commonly used pie chart visualization. In the following, we will give examples for some of the evaluated interfaces; for a complete description, see Appendix A.

Table 2 gives an overview of the ten interfaces[4] and introduces their short names used in the rest of the paper. Regarding their characteristics, we differentiate between personalized and non-personalized ones as well as between approaches that rely on content-information about the recommended items and those who do not[5].

Rank	**Interface long name**	**Interface short name**	**Personalized**	**Content data**
1	Histogram with grouping	`clusteredbarchart`	yes	no
3	Neighbor ratings histogram	`barchart`	yes	no
4	Table of neighbors rating	`neighborsrating`	yes	no
7	MovieLens percent confidence in prediction	`confidence`	yes	no
10	# neighbors	`neighborscount`	yes	no
15	Overall percent rated 4+	`rated4+`	no	no
21	Overall average rating	`average`	no	no
n/a	Pie chart	`piechart`	yes	no
n/a	Tag cloud	`tagcloud`	no	yes
n/a	Personalized tag cloud	`perstagcloud`	yes	yes

Table 2: Explanation interfaces evaluated in our study along with their corresponding ranking according to the study by Herlocker et al. [2000].

3.1. Herlocker et al.'s explanation methods

In [Herlocker et al., 2000], twenty-one different explanation interfaces were compared. In their study, they considered *persuasiveness* as the only quality dimension and determined a ranked list of the best-performing interfaces in this respect, see the "Rank" column in Table 2. The ranked list was then organized in three different groups based on the statistically significance of the observed differences.

[4] We use the terms explanation interface and explanation type interchangeably in this work.

[5] It is important to know that in this study we view content and the visualization to be tightly interrelated in explanation interfaces as done in previous work [Herlocker et al., 2000; Bilgic and Mooney, 2005; Vig et al., 2009; Gedikli et al., 2011] and do not evaluate effects of content and visualization separately.

Since the goal of our study is to analyze the effects of explanation types in several dimensions, we did not simply select the best-performing interfaces from their study but picked interfaces from all three groups. The reason for this choice is that interfaces which perform well in one dimension can perform poorly in other dimensions and vice versa. Examples for such trade-offs are "effectiveness vs. satisfaction" or "effectiveness vs. persuasiveness" as reported in [Tintarev and Masthoff, 2012] and [Gedikli et al., 2011].

In our study, we aimed to cover a wide range of explanation types. To keep the experiments manageable, we limited ourselves to a spectrum of interfaces covering a variety of different types. We selected the explanation types from Herlocker et al.'s study as follows.

- From the top of their list we included the explanation type called *histogram with grouping*, which performed best in their study (see Figure 1), the *neighbor ratings histogram* (ranked 3^{rd}), and *table of neighbors rating* (ranked 4^{th}). As mentioned above, we also included a pie chart based interface which represents a pie chart visualization of the same data presented in the *table of neighbors rating* interface.

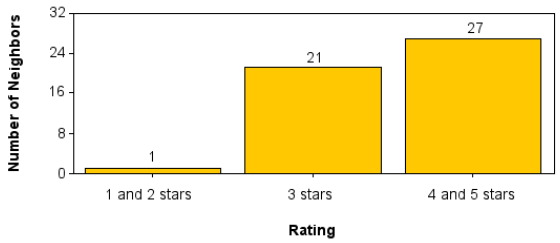


Figure 1: The *histogram with grouping* interface, which performed best in the study of Herlocker et al. [2000]. The x-axis shows the different rating categories (low, medium, high); on the y-axis, the number of neighbors who gave a specific rating is given.

- From the middle block of their list we selected the explanation types *MovieLens percent confidence in prediction*, *# neighbors*, and *overall percent rated* 4+, which were ranked 7^{th}, 10^{th}, and 15^{th} respectively.
- From the end of their list, we picked the interface *overall average rating* (ranked 21^{st}), which performed worst according to the quality dimension persuasiveness. We decided to include the interface due to its popularity on large-scale Web sites, see Figure 2.

Figure 2: IMDb's popular *overall average rating* interface, which performed worst in the study of Herlocker et al. [2000].

3.2. Tag cloud based explanations

Tag clouds were introduced in [Gedikli et al., 2011] as another way of visualizing explanations in recommender systems. Similar to the approach in [Vig et al., 2009], the basic assumption is that each recommendable item can be characterized by a set of tags or keywords. These are

provided by the user community (e.g., of a movie portal) or are automatically extracted from some external source.

- Non-personalized tag clouds

  Figure 3 shows an example of a non-personalized tag cloud used for explanations. The tag cloud contains a set of user-provided keywords (tags) that are relevant for a recommended item. Within the tag cloud, the keywords are arranged in alphabetical order. Larger font sizes indicate higher relevance of a term, which is determined by the number of times a tag was applied to an item.

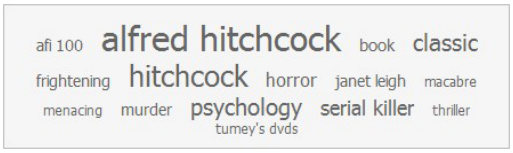


Figure 3: Non-personalized tag cloud.

- Personalized tag clouds

  The approach of using personalized tag clouds (`perstagcloud`) is based on the concept of *item-specific tag preferences* as introduced in [Vig et al., 2010] and [Gedikli and Jannach, 2010]. The difference to the non-personalized version is that the visualization integrates the additional information if a user has a positive (encoded in blue color), negative (red) or neutral (grey) sentiment towards the concept behind each tag. Figure 4 shows an example of a personalized tag cloud.

Figure 4: Personalized tag cloud.

  Another feature of the `perstagcloud` is that when the user inspects a tag cloud and moves the mouse over a specific tag, we display additional information about the items (e.g. movies) that have been attached this tag. This idea was inspired by the keyword-style explanation (KSE) interface proposed by Bilgic and Mooney [2005]. In the experiments, we did however not include KSE into our comparison as we could already show in our previous work that tag clouds were more effective than the keyword-style approach.

  In order to create personalized, tag-based explanations, detailed information about the user's sentiment toward individual tags is required. Acquiring such preferences directly from users as done in our previous study [Gedikli et al., 2011] is however comparably tedious for the user. In this work, we therefore propose a method to *estimate* the user's tag preference in an automated process based on his overall preferences for an item and the tags that are attached to it. In particular, we propose a neighborhood-based scheme similar to the one used in [Gedikli and Jannach, 2013], which determines a preference estimate that is both user- and item-dependent. Details about the procedure are given in Table 3.

Estimating user- and item-specific tag preferences. We propose to use the function $\hat{r}_{u,i,t}$ shown in Equation (1), which we also used in [Gedikli and Jannach, 2013] to generate more accurate rating predictions based on tag preferences.

$$\hat{r}_{u,i,t} = \frac{\sum_{m \in similarItems(i,I_t,k)} w(m,t) * w_{rpa}(r_{u,m}) * \mathcal{R}(r_{u,m})}{\sum_{m \in similarItems(i,I_t,k)} w(m,t) * w_{rpa}(r_{u,m})} \tag{1}$$

The function is user- and item-dependent and returns an estimated preference value for a tag t given a user-item pair (u, i). The main idea is to consider both the overall ratings of items that are similar to the target item i as well as the relative importance of individual tags. Given a set I_t of items tagged with tag t, the function $similarItems(i, I_t, k)$ returns the collection k of the most similar items to i in I_t. The similarity of items is calculated using the adjusted cosine similarity metric. The weight $w(m, t)$ represents the relevance of a tag t for an item m. We use the following simple metric to determine a tag's relevance value, which assigns more weight to tags that have been used more frequently by the user community to characterize an item.

$$w(m,t) = \frac{number\ of\ times\ tag\ t\ was\ applied\ to\ item\ m}{overall\ number\ of\ tags\ applied\ to\ item\ m} \tag{2}$$

When relying on the user's explicit overall rating $r_{u,m}$, no prediction can be made for a tag preference if user u did not rate any item m tagged with t, i.e., if $I_t \cap ratedItems(u) = \emptyset$. We therefore apply the recursive prediction strategy as described in [Zhang and Pu, 2007] and first calculate a prediction for $r_{u,m}$, in case this rating is not available. The function $\mathcal{R}(r_{u,m})$ either returns $r_{u,m}$ if such a rating exists or the estimated value $\hat{r}_{u,m}$. An additional weight value $w_{rpa}(r_{u,m})$ is applied to the recursively predicted value $\hat{r}_{u,m}$ where $w_{rpa}(r_{u,m})$ is defined as follows:

$$w_{rpa}(r_{u,m}) = \begin{cases} 1, & r_{u,m} \text{ is given} \\ \lambda & r_{u,m} \text{ is not given} \end{cases} \tag{3}$$

The combination weight threshold λ is a value between $[0, 1]$. In our study, the parameter λ was set at 0.5 as suggested in [Zhang and Pu, 2007] as an optimal value. We empirically determined $k = 50$ as a suitable value for the neighborhood-size k in Equation (1).

In order to classify tags as positive, negative, and neutral, we used the user's average rating value $\overline{r_u}$ to divide the existing tags into two lists, where one list contains tags whose estimated tag preference is above the user's average and another that contains those which are rated below the average. The tags in each list are then sorted by their predicted preference value in ascending order. We then classify the tags in the lower quartile $Q_{.25}$ of the first list as negative and in the upper quartile $Q_{.75}$ of the second list as positive. All the other tags are classified as neutral. This classification is finally used when generating the personalized tag clouds – positive tags are printed in blue, negative tags are printed in red.

Other tag preference inference schemes are of course possible and might help to further improve the accuracy of the personalized tag clouds. According to [Sen et al., 2009], approaches based on the standard TF-IDF weighting scheme however do not seem to work well in this domain.

Table 3: Procedure to estimate tag preferences.

4. Experimental setup

We adopted a two-phase approach to analyze the effects of different explanation types on users. In the first phase, we conducted a laboratory study with 105 subjects. During this study, the participants interacted with a movie recommendation system and were shown different types of explanations as described in the previous section. The data used later on in the analysis consisted both of the usage logs as well as of the questionnaires that were filled out by the participants at the end of the study. In the second phase, we conducted qualitative, semi-structured interviews with 20 participants that helped us interpreting and validating the results of the quantitative study.

4.1. Setup of the quantitative study

4.1.1. Experimental procedure

The procedure which we used for evaluating the different explanation interfaces in the laboratory study is based on the evaluation protocol proposed in [Bilgic and Mooney, 2005] :

Procedure 1 Experimental procedure used in the laboratory study

1: Get sample ratings from the user.
2: **R** = Set of recommendations for the user.
3: **E** = Set of explanation interfaces.
4: **for all** randomly chosen (r, e) in **R** x **E** **do**
5: Present explanation using interface e for recommendation r to the user.
6: Ask the user to rate r and measure the time taken by the user.
7: **end for**
8: **for all** recommendation r in **R** **do**
9: Show detailed information about r to the user.
10: Ask the user to rate r again.
11: **end for**
12: Ask the user to rate the explanation interfaces.

The procedure consists of the following main phases:

1. *Preference acquisition* (line 1): At the beginning, the participants were asked to provide overall ratings for at least 15 items from a collection of about 1,000 movies.

2. *Generating personalized recommendations* (line 2): The personalized movie recommendations **R** were computed using a user-based nearest-neighbor collaborative filtering algorithm with a neighborhood-size of 50.

3. *Assessing items based on explanations* (lines 2-7): Users were asked to estimate for a number of recommended movies if they will like it nor not given only the explanation generated by the system[6]. The chosen evaluation interface e and movie recommendation r was randomly selected in order to minimize the effects of seeing recommendations or interfaces in a special order. If users guessed which movie they were rating based on the explanation, they were instructed to skip it.

[6] Also the movie titles have been hidden as they might carry explanatory information.

4. *Re-assessing items based on additional information* (lines 8-11): In this phase, we presented the recommended movies another time to the user and asked them to rate the movies. This time, however, we disclosed the movie title and as well as detailed information about the recommended movies such as a trailer, the cast overview, the storyline, plot keywords, and the corresponding genres.

5. *Questionnaire on satisfaction and transparency* (line 12): At the end of the study, the participants were asked to rate their *satisfaction* with the different explanation types on a 1 (lowest) to 7 (highest) rating scale. Furthermore, we measured the user-perceived level of *transparency* for each interface using the same scale. The order in which the different interfaces were presented for rating was again randomized[7].

Beside asking for explicit feedback on satisfaction and transparency at the end, the experimental protocol helped us to measure efficiency, effectiveness and direction of persuasiveness as follows.

- *Efficiency*: In phase 3, we measured the time needed by the user to assess an item based only on the provided explanations.

- *Effectiveness*: We measured *effectiveness* by calculating the differences between ratings provided in the lines 6 (explanation-based rating) and 10 (information-based rating). If both ratings coincide, we have a perfect match with a difference of 0 and the explanation helps the user to accurately estimate the quality of an item.

- *Direction of persuasiveness*: This measurement depends on the effectiveness measure. A strong difference between the explanation-based and information-based ratings means low effectiveness but high persuasiveness and vice versa. If the explanation-based rating is higher than the information-based one, the explanation interface leads to positive persuasiveness and negative persuasiveness in the other case.

The study was conducted in one of the computer labs of our university where the students interacted with a Web-based software application that implemented the required steps of the protocol. The participants completed the experiment one by one and were accompanied by an assistant, who observed the experiment and could help in case of technical questions.

We parameterized our software in a way that every participant was confronted with each of the 10 explanation interfaces **E**. Each interface type was used to explain at least 3 different movie recommendations **R** for the user. Therefore, each participant was asked to provide at least 30 ratings based on the explanations alone. In order to perform the statistical tests properly, the explanation-based ratings from each user were averaged for each explanation type. Detailed statistics about the actual number of ratings will be given later in Section 4.1.4.

4.1.2. Evaluation system and data set details

We developed a software application, with which the study participants interacted and which implemented all the necessary steps to conduct the described experiment. Using this software tool and the defined environment in our computer lab, we could keep possible influence factors such as the effects of differing hard- or software under control.

[7]Details of the questionnaire items are provided in Appendix B.

We have chosen movies as an application domain for the experiments, in particular because of the availability of a public data set which, beside ratings, also contains user-provided tags for the movies. Specifically, we used a subset of the "MovieLens 10M Ratings, 100k Tags" data set[8]. The data set contains about 10 million ratings, about 100,000 tags applied to more than 10,000 movies by more than 70,000 users of the online movie recommender service MovieLens. However, no explicit tag preferences were available and we used the scheme described in Table 3 to estimate the individual tag preferences.

The limited quality of user-contributed tags is one of the major issues when developing tag-based recommendation approaches. In [Sen et al., 2007], it was for example observed that only 21% of the tags in the MovieLens system had adequate quality to be displayed to the user. We therefore selected a subset of the original 10M MovieLens data set because some explanation types require a minimum amount of data with sufficient quality.

In order to improve the quality of the data set, we applied the following filter operations and constraints on the 10M MovieLens data set in order to delete tags, users, or items for which not enough data was available.

- We removed stop-words from the tags such as "a", "by", "about", and "so", by applying stemming [Porter, 1997] and by filtering out noisy tags, which contain a certain number of characters that are not letters, numbers or spaces, e.g., elements such as smileys.
- We required that a tag has been applied by at least 2 different users. This approach was also followed in previous work. In [Vig et al., 2009], for example, the authors require that "a tag has been applied by at least 5 different users and to at least 2 different items".
- We further pre-processed the data by removing items with less than 100 ratings and less than 10 tags, of which at least 7 must be different.

Table 4 shows the MovieLens 10M data set characteristics before and after the application of the filter operations and constraints.

Data Set	#Ratings	#Users	#Items	#Tags
MovieLens 10M	10,000,054	71,567	10,681	95,580
MovieLens 10M Subset	5,597,287	69,876	963	44,864

Table 4: Data set characteristics. We used a subset of the MovieLens 10M data set in our experiment.

4.1.3. Participants of the laboratory study

We recruited 105 participants from ten different countries for our study. In order not to focus on students at our institution alone, we tried to include participants from different demographic groups. Still, most users were generally interested in movies and had at least high-school education. With respect to their cultural background, more than half of the participants were Germans (52%). Another 34% were residents of Germany with Turkish origin (some of them with German citizenship). The German language skills of all participants were sufficient to participate in the experiment and to understand the explanations and the questionnaire. The average age was 28 years. More details about the participants are given in Appendix C.1.

[8] `http://www.grouplens.org/node/73`

4.1.4. Statistics about the data collected in the experiment

In the initial training phase of the experiment, the participants provided 2,370 overall movie ratings (about 23 ratings on average). Figure 5 shows the distribution of the ratings. Users in general preferred to rate movies they liked, i.e., a *positivity bias* can be observed which is in line with the findings of previous work [Marlin et al., 2007; Vig et al., 2010].

In the explanation-based rating phase, we collected 3,108 movie ratings based only on the explanations (about 30 per user on average). In only 1% of the cases (42 out of 3,150), the users recognized the movie and skipped to the next explanation without providing a rating. Note that users were allowed to provide as many ratings as they wanted.

Finally, during the information-based rating phase, users provided 315 movie ratings based on more detailed information (3 ratings per user on average).

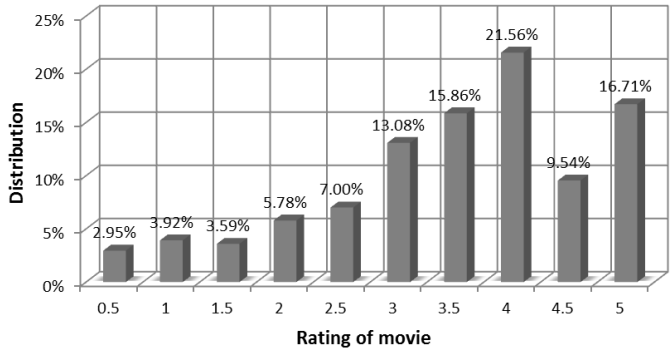


Figure 5: Distribution of movie ratings on a 5-point rating scale with half-point increments.

4.2. Design of the interview-based study

One of the outcomes of the analysis of the laboratory study described in the previous section is a set of guidelines regarding the design of explanation interfaces for recommender system. In order to cross-check the validity of these guidelines, we conducted an in-depth interview-based study which complements the statistical analysis of the lab experiment.

In the literature, such semi-structured interviews are reported to be effective for working with small samples and to be helpful for supplementing and validating information derived from other sources [Laforest et al., 2009]. The qualitative data collected in these interviews should thus help us to obtain more evidence that our guidelines are plausible and reliable.

For the interviews, we recruited 20 subjects who had not participated in the experiment described in the previous sections. The average age was 25 years and most of them were computer science students at the university. The details of the interview plan are provided in Appendix D.

The procedure was as follows. At the beginning of the interview, each participant was briefly introduced to the field of recommender systems. Next, we asked the participants to describe in their own words what they think what the role of explanations could be in the context of recommender systems and how explanations may look like. In order to avoid any bias, we did not provide descriptions, definitions, or even examples for explanations in the introduction to the field.

Then, we conducted semi-structured interviews in which we asked the participants about their opinions towards explanations and asked questions such as *"Is it important to you to understand*

the reasons why a particular item was recommended?" or *"What could be the possible goals of providing explanations for recommendations?"*.

5. Observations and design guidelines

In this section, we will both discuss the observations made in the laboratory study as well as our findings from the interview-based validation study. Based on these insights, we derive a set of possible guidelines for the design of explanation interfaces for recommendation systems.

We will structure our discussions regarding the experimental study according to a number of more general and some specific research questions, to which our work shall contribute.

Efficiency	*Are there some explanation interfaces that help users assess an item more quickly?* *Do explanation interfaces based on content information (tag clouds) require significantly more effort by users than others?*
Effectiveness and Persuasiveness	*Which explanation interfaces are the most effective ones?* *Does personalization help to increase the effectiveness of the content-based method?* *Do some explanation interfaces lead to (potentially undesired) effects of overestimate or underestimate-oriented persuasiveness?*
Transparency and satisfaction	*Which explanation interfaces are considered to be transparent by the users?* *How is the overall satisfaction of users for each explanation type?*
Dependencies and trade-offs	*How are the different factors related?* *Which factors have an impact on the users' overall satisfaction with the explanation interface?*

Table 5: Summary of research questions to be addressed.

5.1. Methods and tools for statistical analysis

SPSS 20 was used for data analysis and all the tests were done at a 95% confidence level. Regarding the statistical methods applied in our analysis, we used the Friedman test throughout this work to analyze whether observed differences between the various explanation types are statistically significant. We also used t-tests and repeated-measures ANOVA, but we will only report the results of the non-parametric Friedman test here since the results were similar across all tests. The Friedman test is suggested by Demšar [2006] for a comparison of more than two systems.

Once a significant result was found, we applied the post-hoc Wilcoxon Signed-Rank test to identify where the specific differences lie. In order to interpret our Wilcoxon test result, a Bonferroni correction was accordingly applied and thus all the effects are reported at a $p < 0.005$ level of significance, if not stated otherwise. Detailed test statistics are listed in Appendix C.

5.2. Observations and measurements for the laboratory study

5.2.1. Efficiency of explanations

Recall that efficiency stands for the ability of an explanation to help the users make their decisions faster. We defined the decision time in our study as the time required by the users to submit a rating after seeing the corresponding explanation interface. Figure 6 shows the mean time (in seconds) needed by the users for each explanation type together with the standard deviation. The figures in parentheses indicate the group of Herlocker's methods an individual explanation type belongs to. The groups are ranked according to how well their methods performed with respect to persuasiveness.

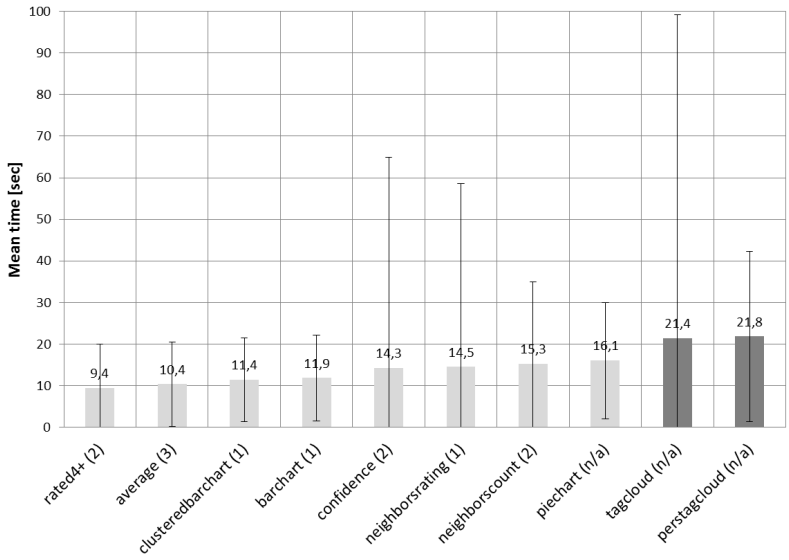


Figure 6: Mean time for submitting a rating after seeing the corresponding explanation type (**efficiency**). Light gray bars indicate explanations with a mean time significantly different from the base case `perstagcloud` ($p < 0.005$, $N = 291$). The figures in parentheses indicate to which of Herlocker et al.'s groups the method belongs to.

The results show that that users interacting with the content-based tag cloud explanations in fact need significantly more time for decision making. When presented with the most efficient interfaces (`rated4+` and `average`), users needed less than half of the time. This observation is not particularly surprising, given that the tag cloud interface conveys more information for the decision process than the simple average rating display. Users that were confronted with the personalized version of the tag clouds on average needed more time. The additional time that is theoretically needed to interpret the color codings was however small and the differences were not statistically significant.

Modest differences can be observed between Herlocker et al.'s explanation approaches. When considering their ranking with respect to persuasiveness in combination with the required decision times, no clear trend can be observed except that the non-personalized methods `rated4+` and `average` were slightly more efficient than the others.

Overall, the data clearly shows that some forms of explanations help users decide more quickly than others. The interesting question addressed later on however is whether and to which

extent the efficiency of an explanation type is related to decision quality and finally user satisfaction.

5.2.2. Effectiveness and persuasiveness

Effectiveness is the ability of an explanation to help a user to accurately assess the quality of the recommended item and make better decisions. We approximate the effectiveness of an explanation type by considering the correlation and standard deviation values between explanation-based and information-based ratings (see Section 4.1). Figure 7 shows the mean difference values for the explanation types together with the standard deviation in ascending order of the magnitude of the difference.

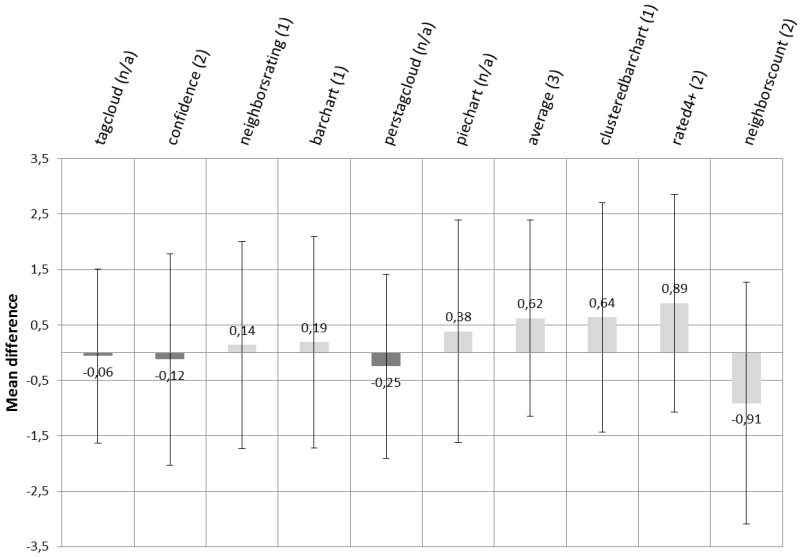


Figure 7: Mean difference of explanation-based ratings and ratings given based on more detailed information (**effectiveness**). Light gray bars indicate explanation types with a mean difference significantly different from the base case `perstagcloud` ($p < 0.005$, $N = 291$). The figures in parentheses indicate to which of Herlocker et al.'s groups the method belongs to.

The results indicate that the `tagcloud` interface leads to a very small difference between the explanation-based rating and the rating based on detailed information, which can be interpreted as a sign of good effectiveness and low persuasiveness. The `neighborscount` interface represents the other extreme and leads to the effect that users strongly underestimate how they would like a movie when seeing only how many like-minded users (neighbors) rated the movie. Thus, the probably undesired effect of persuasiveness is high.

However, only observing a mean value close to 0 might not be sufficient to tell whether an explanation is effective or not. Consider an explanation which always leads to the effect that users give ratings which are close to the overall mean rating. In this case, the mean difference between explanation ratings and actual ratings can also be 0.

To reduce the risk of averaging effects, we analyzed the correlation of the explanation-based ratings and the information-based ratings as well as the corresponding standard deviation values

as proposed by Bilgic and Mooney [2005]. Overall, from the user's point of view, a good explanation interface has (a) a mean difference value of 0, (b) a high positive correlation between explanation ratings and information-based ratings, and (c) a low standard deviation value.

We show the correlation and standard deviation values in Table 6. The triangles in the subsequent tables indicate how the tables are sorted. In Table 6, the data rows are sorted in descending order of the correlation value.

#		Pearson Corr ▽	Std Dev	Sig Level	Herlocker Group
1	`tagcloud`	**0.506**	1.570	0.000	-
2	`perstagcloud`	**0.504**	1.661	0.000	-
3	`confidence`	**0.243**	1.904	0.000	2
4	`piechart`	**0.239**	2.002	0.000	-
5	`neighborsrating`	**0.231**	1.867	0.000	1
6	`average`	**0.226**	1.772	0.000	3
7	`barchart`	**0.193**	1.903	0.001	1
8	`clusteredbarchart`	**0.192**	2.071	0.001	1
9	`rated4+`	-0.049	1.960	0.408	2
10	`neighborscount`	-0.053	2.182	0.367	2

Table 6: Pearson correlation values between explanation ratings and actual ratings and standard deviation values of the mean differences. Correlation is significant at the 0.01 level (2-tailed).

Figure 7 shows that the mean differences for the tag cloud explanations are close to 0, which – as mentioned above – is an indication that they help users accurately estimate the quality of an item. Moreover, from Table 6, we see that the tag cloud interfaces in addition have the highest correlation with the lowest standard deviation. The interfaces `confidence`, `neighborsrating`, and `barchart` also have mean difference values close to 0. However, `perstagcloud` has a correlation which is twice as high, and at the same time it has a lower standard deviation.

Contrary to our intuition, the results indicate that the explanations based on `tagcloud` are slightly more effective than their personalized counterpart (`perstagcloud`). A similar phenomenon was reported by Tintarev and Masthoff [2012], who also observed that personalization was detrimental to effectiveness. In their work, the authors report results from three experiments in which their personalization method hindered effectiveness but increased the satisfaction with the explanations[9]. One possible reason that the personalization did not lead to higher effectiveness in our study might lie in the fact that our method to estimate the tag preferences automatically was not accurate enough given the comparably small amount of tag data. This in turn could lead to more polarized opinions, comparable to the work of Tintarev and Masthoff [2012]. For example, a personalized tag cloud explanation may have biased users to expect a movie that they will really like because their favorite actors were mentioned in positive blue tags. However, they might be later on disappointed when those actors only had a minor role.

Regarding the aspect of persuasiveness, most of the explanation interfaces from Herlocker et al., and in particular the non-personalized ones, cause the users to overestimate the true value of a movie. Again, no clear correlation with the persuasiveness-based ranking of their study can be observed. Interestingly, the `neighborscount` is the only one of their methods that lead to an underestimated-oriented persuasiveness. However, the difference is not significant.

[9] We will see later on that this is also the case for the tag cloud explanations.

Overall, the results show that the newly proposed and content-based tag cloud interfaces are among the most effective explanation types which at the same time have a very small tendency to cause the user to over- or underestimate the real value of a recommended item. Given that the tag cloud method is the only content-based technique in this comparison, we consider this as being an indication that users are able to evaluate explanation types based on content information more precisely. This is also in line with the findings of the study of Herlocker et al. [2000], where the content-based method "Favorite actor or actress" interface was among the best-performing ones.

In summary, our first suggestion when designing an explanation interface is therefore:

Guideline 1 *Use domain specific content data to boost effectiveness.*

Previous studies did not differentiate between explanation types that use content information and those that do not. When we designed our study we did not particularly focus on this aspect either and only included one type of explanations that uses content information, in particular as we could already observe in a previous study that tag cloud explanations are favorable when compared with a keyword-style approach. Since content information however seems to play a crucial role, a more systematic analysis of the effects of incorporating different amounts and types of additional data into the explanations represents an important direction for future research.

5.2.3. User-perceived transparency

We measured the *user-perceived* level of transparency by directly asking the users whether the provided explanation types helped them understand how the recommender system works. Users could rate each explanation interface on a scale from 1 (not transparent at all) to 7 (highest transparency). The results are shown in Table 7.

#		transparency ▽	Std Dev	Herlocker Group
1	`perstagcloud`	5.61	1.53	-
2	`barchart`	5.51	1.26	1
3	`piechart`	5.41	1.21	-
4	`clusteredbarchart`	5.40	1.25	1
5	`rated4+`	5.27	1.28	2
6	`neighborsrating`	5.12	1.13	1
7	`average`	5.07	1.46	3
8	`tagcloud`	**5.05**	1.60	-
9	`confidence`	**4.65**	1.34	2
10	`neighborscount`	**2.80**	1.70	2

Table 7: Mean rating of the users for each explanation type regarding **transparency** using a 1-to-7 scale. Numbers printed in bold face indicate the explanation type with a mean rating that is significantly different from the base case `perstagcloud` ($p < 0.005$, $N = 105$).

The novel `perstagcloud` interface was perceived by the users to be the most transparent one, followed quite closely by the three chart-based interfaces. The differences between these explanation types were however only modest and two of the best-performing chart-based methods of Herlocker et al.'s study also led to comparably high levels of user-perceived transparency.

The `neighborscount` interface can be found at the other end of the list and was evaluated much worse by the users than the others. This explanation type merely consists of a statement

about how many similar users have rated the movie. Users probably did not understand how the number of neighbors alone was sufficient for the system to create a personalized recommendation.

The question whether or not providing content information is generally well-suited to increase the transparency of recommendations, cannot be conclusively answered given our observations. While the `perstagcloud` was considered to be the most transparent explanation form, the perceived transparency of its non-personalized version (`tagcloud`) was significantly lower.

5.2.4. Satisfaction

In our questionnaire at the end of the laboratory study we explicitly asked users about their overall satisfaction with the explanation interfaces. Table 8 shows the participants' mean rating for each explanation interface using again a scale from 1 to 7.

#		satisfaction ▽	Std Dev	Herlocker Group
1	`perstagcloud`	4.96	1.93	-
2	`average`	4.70	1.39	3
3	`rated4+`	4.63	1.50	2
4	`tagcloud`	4.59	1.91	-
5	`clusteredbarchart`	4.57	1.60	1
6	`barchart`	4.56	1.40	1
7	`confidence`	4.45	1.39	2
8	`piechart`	4.32	1.75	-
9	`neighborsrating`	**3.95**	1.46	1
10	`neighborscount`	**2.09**	1.38	2

Table 8: Mean rating of the users to each explanation interface regarding **satisfaction** on a 1-to-7 rating scale. Numbers printed in bold face indicate the explanation interfaces with a mean rating that is significantly different from the base case `perstagcloud` ($p < 0.005$, $N = 105$).

The novel `perstagcloud` interface led to the overall highest average satisfaction. The differences between the best-accepted method and those ranked next are however modest. Statistically significant differences between the `perstagcloud` and other methods could only be observed for the `neighborsrating` and `neighborscount` explanation types.

Interestingly, the non-personalized explanation types `average` and `rated4+` were well accepted by the users, even though they are very simple and merely present an item's average rating or the percentage of users who liked the item, respectively. Note that in the study of Herlocker et al., the `average` interface actually performed worst, while it is second in our study. Remember, however, that their study was conducted from a promotion perspective (persuasiveness) only. Our study thus provides additional evidence that it is important to evaluate explanation interfaces from different points of view in parallel as suggested in [Tintarev and Masthoff, 2012].

A further reason for the comparably high satisfaction of users with the non-personalized interfaces `average` and `rated4+` could lie in their simplicity and in the fact that these types of information can nowadays be found on many popular (shopping) web sites. Thus, users are probably acquainted to rely on such information in their decision process.

The acquaintance of users with the visual appearance of tag clouds was also one of the key factors for us to use tag clouds in the explanation process as nowadays, such tag clouds can be

found on a number of popular Social Web sites such as Delicious and Flickr[10] where they are used as a visualization and interaction technique[11]. Overall, we see the result regarding satisfaction as an indication that the combination of various features in the `perstagcloud` approach (including personalization, the provision of content-information and the usage of a well-known visualization approach) can lead to improved user satisfaction despite the fact that the explanation type demands higher cognitive effort by the users when compared to more simple approaches. Based on these observations, we suggest:

Guideline 2 *Use explanation concepts the user is already familiar with, as they require less cognitive effort and are preferred by the users.*

5.2.5. Relationships between variables

In order to obtain a deeper understanding of the relationships between the different quality factors discussed so far, we conducted a *Path Analysis* which shall help us to characterize possible dependencies in a quantitative manner.

The analysis was done using regression analysis, which is a standard approach in social sciences and in educational research to study potential causal relationships [Tuijnman and Keeves, 1994]. SPSS AMOS 20 was used for model building and the actual analysis step.

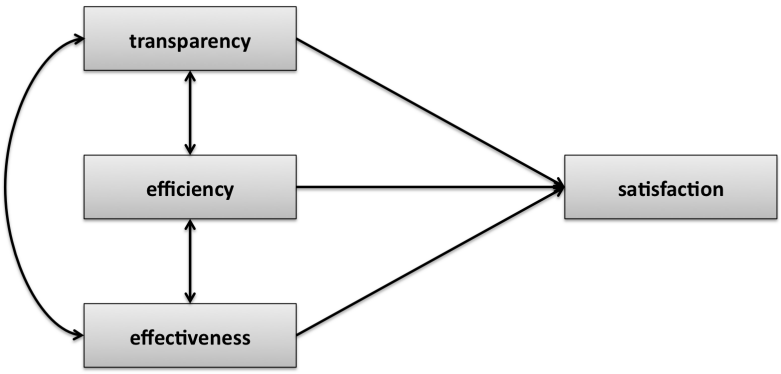


Figure 8: The SPSS AMOS path analysis model showing the dependencies between the variables.

Figure 8 shows the model which we used as input for the analysis. Each quality factor is represented by an independent or dependent variable in the model. We see satisfaction as the dependent variable which can be influenced by the quality factors efficiency, effectiveness, and user-perceived transparency, which are consequently modeled as independent variables. The edges between the independent variables represent the covariance parameters to estimate. We conducted a path analysis for each explanation interface separately. The results of each analysis run are shown in Table 9 and Table 10. The rows of the table are sorted according the interfaces' ranking with respect to satisfaction (Table 8).

Table 9 shows the maximum likelihood estimates of the regression weights, along with their accompanying R squared values, indicating the goodness of fit for estimating the parameters.

[10] `http://www.delicious.com`, `http://www.flickr.com`

[11] Some of the participants of our experiment had not seen tag clouds before, which we see as the main reason for their comparably high standard deviations.

	transparency	efficiency	effectiveness	R^2
	↓			
	satisfaction			
perstagcloud	**0.883**	0.002	0.238	0.505
average	**0.482**	0.010	0.235	0.339
rated4+	**0.674**	0.016	**0.302**	0.469
tagcloud	**0.888**	0.003	0.237	0.510
clusteredbarchart	**0.760**	-0.003	0.108	0.350
barchart	**0.773**	-0.005	0.144	0.467
confidence	**0.509**	-0.002	0.066	0.254
piechart	**0.705**	-0.011	0.317	0.330
neighborsrating	**0.547**	0.001	0.112	0.187
neighborscount	**0.473**	0.008	0.171	0.467
∅	0.669	0.002	0.193	0.388

Table 9: Maximum likelihood estimates of the regression weights. Bold figures indicate weights with a significant effect ($p < 0.001$, $N = 291$).

	transparency		transparency		efficiency	
	↕		↕		↕	
	effectiveness		efficiency		effectiveness	
	Cov	Corr	Cov	Corr	Cov	Corr
perstagcloud	-0.054	-0.032	-1.741	-0.090	0.601	0.044
average	0.326	0.195	0.626	0.058	0.194	0.023
rated4+	0.356	0.208	-0.473	-0.059	0.107	0.013
tagcloud	-0.052	-0.031	-1.743	-0.090	0.601	0.044
clusteredbarchart	-0.187	-0.123	-0.415	-0.054	-0.508	-0.067
barchart	-0.338	-0.217	-0.562	-0.072	0.241	0.031
confidence	0.249	0.145	-0.590	-0.016	0.950	0.026
piechart	0.162	0.101	-0.825	-0.082	-0.622	-0.055
neighborsrating	-0.014	-0.010	-1.520	-0.560	-0.622	-0.021
neighborscount	0.820	0.292	0.924	0.044	4.877	0.238
∅	0.127	0.053	-0.632	-0.092	0.582	0.028

Table 10: Covariances and correlations among the independent variables.

Effects of transparency on satisfaction. The results clearly show that user-perceived transparency – independent of the used interface – has a significant positive effect on user satisfaction. Because both transparency and satisfaction use a 7-point Likert scale (from "not at all" to "a lot"), we have also considered the Pearson correlation coefficient and the Spearman rank correlation coefficient between transparency and satisfaction which are 0.84 and 0.57 respectively. We see this as a strong indication that users are generally more satisfied with explanation facilities which provide justifications for the recommendations. This even holds when the user's beliefs of how the system works are actually wrong. Our design suggestion is therefore:

Guideline 3 *Increase transparency through explanations for higher user satisfaction.*

Effects of efficiency on satisfaction. When we look at the path analysis results for efficiency, we can observe that decision time – in contrast to transparency – seems to have no influence on user

satisfaction. The average regression weight of efficiency is close to 0. Based on this observation and the findings reported above, we derive a further design suggestion for explanation interfaces:

Guideline 4 *Explanation types should not primarily be optimized for efficiency. Users take their time for making good decisions and are willing to spend the time on analyzing the explanations.*

This guideline in general does not rule out that there might be other explanation types not considered in our study which are so easily and quickly understandable by users such that there exists also a positive effect on satisfaction. We could however not observe such an effect for the explanation methods used in our study.

Effectiveness and user satisfaction. With respect to the relation of effectiveness and user satisfaction, we see that the average weight for effectiveness is less than one-third of the average weight for transparency. Except for the `rated4+` interface, effectiveness had no significant (short term) effect on user satisfaction in our study.

Nonetheless, we suspect that effectiveness can have a long-term effect on satisfaction which we were not able to capture in our single-session experiment and which would require long-term analysis spanning longer periods of time and multiple sessions.

There are, however, some indications that lead us to this hypothesis that effectiveness is important for the success of a system in the long run. Consider, for example, the comparably simple `average` interface with which the users were generally satisfied. According to the results shown in Figure 7, the interface influences users in a way that they actually overestimate the value or suitability of an item. In the long run, users might therefore be disappointed after experiencing or consuming the recommendation item and sooner or later lose their trust in such simple quality indicators or the recommendation system as a whole. One design advice in this context could therefore be to "enhance effectiveness in order to increase user satisfaction in the long run".

Relationships between independent variables. Table 10 finally contains the covariance and correlation values between the independent variables. As expected, the average correlation values of the independent variables are close to 0, i.e., there is no linear correlation between the variables. The results thus confirm our decision to model efficiency, effectiveness, and transparency as independent variables since we could not detect strong interdependencies among them.

However, for some interfaces, we can observe trade-offs between different dimensions. For example, a low negative correlation value of -0.56 exists between transparency and efficiency for the interface `neighborsrating`, an interface that displays a tabular view of the ratings of the user's neighbors. These observations are in line with those made in previous work. Tintarev and Masthoff [2007b], for example, conclude that high transparency may impede efficiency as the users need more time to inspect an explanation type.

5.3. Results of the interview-based validation study

In the following, we will discuss the observations made in the interview-based study, which we conducted in order to validate the findings of the laboratory study. Given the answers of the 20 participants, we could make the following observations.

5.3.1. The role of explanations in recommender systems

Regarding the function of explanations, the descriptions of the large majority of the participants (16 out of 20) were centered on the topic of transparency. In other words, the majority of the participants saw the primary role of explanations to help users understand how the system works and how the recommendations were generated.

Furthermore, 12 participants stated that it is in general very important to them to understand how the system works. For another 5 participants understanding recommendations was important, but only for unexpected or questionable recommendations. For example, one of them commented that *"I want to know why I get five printer recommendations from Amazon, although I have bought one on Amazon just last week."*. Only 2 participants explained that transparency is less important to them; one had no further comments. Overall, these results support our Guideline 3 on transparency which states that an increased level of transparency of the explanations can help to better meet the users' expectations and to increase the overall user satisfaction.

5.3.2. The role of efficiency

As a next step, we asked the participants about their willingness to invest time to try to understand the system's recommendations. Four participants agreed without any concerns. The majority (80%, 16 out of 20) however said that they would only by willing to invest time under certain circumstances. The specific circumstances mentioned by the participants can be grouped along the following dimensions.

- *Explanation-related aspects*: About 69% (11 out of 16) of the participants mentioned that their willingness to invest more time depends on the quality of the provided explanations, which, for example, should be concise and easy-to-understand.
- *Domain-dependent aspects*: About 25% stated that it depends on whether or not the recommended item is expensive (e.g., a digital camera) or comes with some risk (e.g., an investment product).
- *Recommendation dependent aspects*: Two participants (13%) answered that they would only inspect the explanations in detail when a recommendation is unexpected or questionable. One person stated that explanations were relevant and worth inspecting when they can be used as a basis to influence the provided recommendations.

We then asked these participants if they would be willing to invest more time in case they would expect with some certainty that the explanation would significantly help them to make a good decision. From the 16 participants, 15 agreed. Overall, we see these observations as additional evidence supporting Guideline 4, which states that even though efficiency is important to the user, there can be other aims such as effectiveness which can be even more important. At the same time, the results can also be viewed as indicators supporting the validity of our hypothesis on the relation between effectiveness and long-term satisfaction with the system.

5.3.3. The role of content-related information

Coming back to the first question, when we asked the subjects to define explanations for recommenders in their own words, the answers of 7 participants were in some form related to content data, i.e., they associated explanations with the provision of additional information for the recommended item. 17 participants stated that such information is a prerequisite for making a good decision and 15 of them explicitly declared content information to be the most important

information source for their decisions. On the other hand, 13 participants argued that a social component, such as the opinions and ratings of other users in the system, are also helpful when assessing the quality of a recommendation. This confirms the findings from Herlocker et al. [2000]. However, only 3 participants would prefer this latter type of information over other information sources such as content data. We see these results as an indication for the validity of Guideline 1 which proposes to use domain specific content data to increase effectiveness.

Overall, even though the sample in the interview-based study was of limited size, we see the observations as further evidence supporting the validity of the guidelines presented in this paper.

6. Discussion and conclusions

6.1. Research limitations

Tintarev and Masthoff [2012] argue that the movie domain, which was the focus of this study, suffers from being subjective in nature. However, the results of their study revealed that some of their observations from the movie domain could also be made in the (more objective) domain of digital cameras. Our future work includes an evaluation of whether the effects of explanations reported in this study can also be observed in other domains.

Another aspect to consider in our experimental design is that in our study the (perceived) recommendation quality can in principle vary across users. This, in turn, might have an effect on the user's overall degree of satisfaction with the system. In our study, we therefore restricted the set of available items to a subset of comparably popular movies with at least 100 user ratings to keep this limitation under control. This way, we were able to generate generally well-perceived, high-quality recommendations even when only a limited number of ratings per user was available. Using only relatively popular items might of course have led to a stronger positivity bias when compared with settings in which also unpopular items are recommended.

Finally, in our current study we limited the set of explanations to which we compared our tag cloud approach mainly to a subset of those proposed by Herlocker et al. In our own previous work ([Gedikli et al., 2011]), we have also compared the tag cloud approach with another well-performing explanation method based on content keywords proposed in [Bilgic and Mooney, 2005]. Since our previous study has revealed that tag clouds were better accepted by users than the keyword-style approach, we have not considered this explanation type in the current study. However, we consider a more systematic analysis and comparison of explanations that use different types and amounts of content data to be an important next step of our future work.

6.2. Research outlook: explanations and trust

In previous research, explanations are often discussed in the context of *trust* [Pu and Chen, 2006, 2007]. In our view, the question how explanations help to improve trust remains open to some extent, in particular as trust is a multi-faceted concept which cannot be easily measured directly. Tintarev and Masthoff [2007a] also mention this difficulty of measuring trust and propose measuring it through user loyalty and increased sales. Since we see trust as the long-term relationship between the user and the system, we believe that trust can only be captured in a long-term study.

Here, we will examine trust from a theoretical perspective and show possible directions for future work. Our aim is to discuss the possible influence of the quality factors on trust.

- First, we consider satisfaction as a prerequisite for trust. Satisfaction with the explanations increases the overall satisfaction with the system [Cosley et al., 2003]. A user who is not satisfied with the system is not likely to develop trust in a system. Trust is necessary to keep the user satisfaction sustained over a long period of time.

- Our study showed that user-perceived transparency is a highly important factor for user satisfaction. Thus, we believe that user-perceived transparency is also an important factor for trust, which is also suggested in the literature, see, e.g., Swearingen and Sinha [2002].

- Efficiency, on the other hand, does not appear to be particularly important for user satisfaction. In our study we could not analyze the long-term effects of efficiency.

However, we believe that efficiency has a limited effect on trust. An efficient system might be more comfortable to use since it requires less cognitive effort but may not necessarily be more trustworthy.

Bilgic and Mooney [2005] argue that effectiveness is more important than persuasiveness in the long run as greater effectiveness can help to establish trust and attract users. We could not measure a significant (short term) effect of effectiveness on satisfaction (except for `rated4+`). However, explanations that continuously lead to an overestimate of an item's quality can be risky since the user may get the impression that the recommender system is cheating because it is promoting items without taking the user's true preferences into account. On the other hand, if the explanations continuously lead to an underestimation of the quality, the system may leave the user with the impression that the system generally fails to generate accurate recommendations. Thus, both positive and negative persuasiveness can cause the loss of trust to users.

6.3. Summary

In this work, we have presented the results of a user study in which ten different explanations types were evaluated with respect to several quality factors. Besides explanation types known from the literature, our study included a new visualization method based on tag clouds.

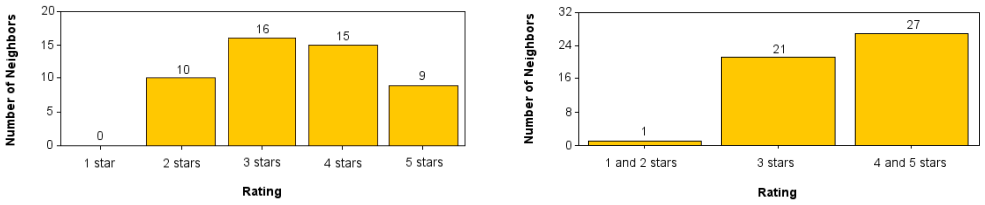
The analysis revealed that explanation types have different effects on users. Some of them for instance help users decide more quickly, whereas others lead to higher levels of perceived transparency and overall satisfaction. A path analysis furthermore revealed a strong relationship between transparency and satisfaction. Overall, the content-based tag cloud explanations were effective and particularly well accepted by users. Furthermore, its personalized variant was helpful to make the underlying recommendation logic transparent for the users. Putting these thoughts together, we hypothesize that in particular the newly proposed tag cloud interfaces are candidates for building trustworthy explanations.

Based on these observations, we derived a set of guidelines of how to build explanations for recommender systems that support the user in the decision making and buying process.

Beside a deeper investigation of the long-term relationship between explanations and trust as described in the previous section, our future work includes the exploration of explanations that are based on explicitly encoded domain knowledge, e.g., in the sense of Zanker [2012].

Appendix A. Explanation interfaces used in the study

Figure A.9 (a) shows a `barchart` explanation which basically contains a standard bar chart histogram showing the distribution of the target user's neighbors' ratings. Figure A.9 (b) is the `clusteredbarchart` explanation, in which the high and low ratings are collapsed to one bar.



(a) `barchart` (b) `clusteredbarchart`

Figure A.9: The `barchart` and `clusteredbarchart` explanations.

Figure A.10 shows the explanation interface `average`, which presents the user with the overall average rating of the target item.

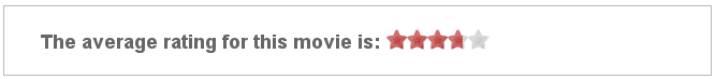


Figure A.10: The `average` interface.

Figure A.11 is the `confidence` interface which corresponds to Herlocker et al.'s *MovieLens percent confidence in prediction* interface.

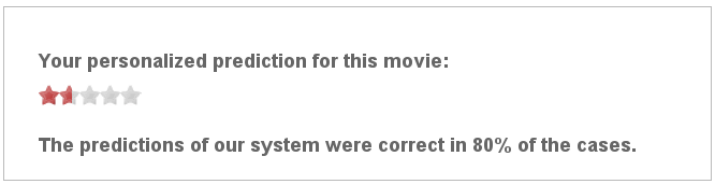


Figure A.11: The `confidence` display.

Figure A.12 is the `neighborsrating` explanation, which is a tabular representation of the ratings within the user's neighborhood. The `piechart` explanation type shown in Figure A.13 represents the same data in a different way.

Beside the new tag cloud interfaces, we finally also included two string-based explanation types from [Herlocker et al., 2000]. The `neighborscount` interface shows the number of neighbors who provided a rating for the target item, whereas the `rated4+` explanation reveals the percentage of ratings for the target item which are equal or greater than 4.

Figure A.12: The `neighborsrating` display.

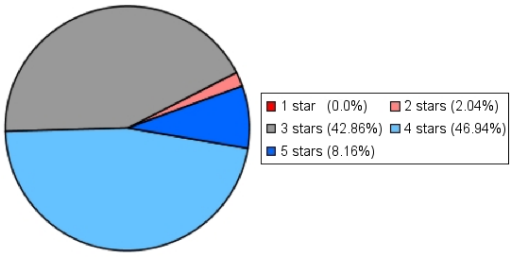


Figure A.13: The `neighborsrating` display.

Appendix B. Translations of questionnaire items for the laboratory study

The study was conducted in the German language. The following questions were asked at the end of the experiment.

*Please rate the explanation interfaces based on the criteria **Ease of Usability** and **Enjoyment**. Also, let us know to which extent the different explanation interfaces were suited to increase the **Transparency** of the system. Transparency means that the explanation interface helps you to understand how the recommendation system works. Please provide a rating on the seven-point scale from "not at all" to "very much".*

You can leave a comment on each explanation interface and tell us what you particularly liked or disliked about it.

Appendix C. Detailed Statistics for Laboratory Study

Appendix C.1. Participants of the laboratory study

Gender	Female	52 (49.52%)
	Male	53 (50.48%)
Education	A-level	58 (55.24%)
	Bachelor	5 (04.76%)
	Master	32 (33.33%)
	PhD	2 (01.90%)
	Other	5 (04.76%)
Nationality	Germany, Turkey, China, Austria, Poland, Bosnia, Czech Republic, Ukraine, Iran, Albania	
Age	19-23	34 (32.38%)
	24-28	29 (27.62%)
	29-33	27 (25.71%)
	34-38	11 (10.48%)
	39-45	4 (03.81%)
Interest in movies	Low	14 (13.33%)
	Normal	48 (45.71%)
	High	43 (40.95%)

Table C.11: Demographic characteristics of participants (total 105).

Appendix C.2. Friedman test

	efficiency	effectiveness	transparency	satisfaction
clusteredbarchart	4.85	6.59	6.26	5.89
barchart	5.04	5.40	6.45	5.93
neighborsrating	5.36	5.38	5.45	4.81
confidence	4.89	4.84	4.62	5.70
neighborscount	5.66	3.23	2.20	2.21
rated4+	3.88	6.98	5.87	5.92
average	4.32	6.57	5.48	6.15
tagcloud	6.86	5.06	5.59	6.00
perstagcloud	7.37	4.83	6.96	6.87
piechart	6.76	6.11	6.13	5.52

Table C.12: Friedman test mean ranks.

	N	Chi-Square	df	Asymp. Sig.
efficiency	291	381.385	9	.000
effectiveness	291	415.036	9	.000
transparency	105	209.494	9	.000
satisfaction	105	180.849	9	.000

Table C.13: Friedman test statistics.

Appendix C.3. Wilcoxon Signed-Rank test

	efficiency		effectiveness		transparency		satisfaction	
	Z	Asymp. Sig.	Z	Asymp. Sig.	Z	Asymp. Sig.	Z	Asymp. Sig.
average-barchart	−2.956[a]	.003	−5.638[a]	.000	−2.600[a]	.009	−.644[a]	.520
average-clusteredbarchart	−2.056[a]	.040	−.059[b]	.953	−2.250[a]	.024	−.502[a]	.616
average-confidence	−2.430[a]	.015	−7.619[a]	.000	−2.442[b]	.015	−1.589[a]	.112
average-neighborscount	−5.577[a]	.000	−11.785[a]	.000	−8.104[b]	.000	−8.427[a]	.000
average-neighborsrating	−4.130[a]	.000	−5.959[a]	.000	−.264[a]	.792	−4.176[a]	.000
average-perstagcloud	−10.345[a]	.000	−7.788[a]	.000	−2.712[a]	.007	−1.290[b]	.197
average-piechart	−8.576[a]	.000	−2.605[a]	.009	−2.270[a]	.023	−1.853[a]	.064
average-rated4+	−1.810[b]	.070	−3.361[b]	.001	−1.809[a]	.070	−.339[a]	.735
average-tagcloud	−8.426[a]	.000	−6.298[a]	.000	−.015[a]	.988	−.275[a]	.784
barchart-clusteredbarchart	−.821[b]	.412	−5.872[b]	.000	−1.049[b]	.294	−.082[a]	.935
barchart-confidence	−.864[b]	.388	−3.310[a]	.001	−4.231[b]	.000	−.778[a]	.436
barchart-neighborscount	−2.850[a]	.004	−9.281[a]	.000	−8.050[b]	.000	−7.747[a]	.000
barchart-neighborsrating	−.911[a]	.362	−.834[a]	.404	−3.132[b]	.002	−3.485[a]	.000
barchart-perstagcloud	−8.253[a]	.000	−3.810[a]	.000	−.988[a]	.323	−2.346[b]	.019
barchart-piechart	−5.880[a]	.000	−3.185[b]	.001	−.728[b]	.466	−1.533[a]	.125
barchart-rated4+	−5.121[b]	.000	−7.123[b]	.000	−1.635[b]	.102	−.088[b]	.930
barchart-tagcloud	−6.666[a]	.000	−2.243[a]	.025	−2.753[b]	.006	−.322[b]	.747
clusteredbarchart-confidence	−.264[b]	.792	−7.417[a]	.000	−4.353[b]	.000	−.476[a]	.634
clusteredbarchart-neighborscount	−3.144[a]	.002	−10.370[a]	.000	−8.109[b]	.000	−7.679[a]	.000
clusteredbarchart-neighborsrating	−1.666[a]	.096	−6.408[a]	.000	−2.102[b]	.036	−3.021[a]	.003
clusteredbarchart-perstagcloud	−9.309[a]	.000	−6.876[a]	.000	−1.568[a]	.117	−1.994[b]	.046
clusteredbarchart-piechart	−6.753[a]	.000	−2.917[a]	.004	−.114[a]	.909	−1.408[a]	.159
clusteredbarchart-rated4+	−4.342[b]	.000	−2.172[b]	.030	−1.024[b]	.306	−.078[b]	.938
clusteredbarchart-tagcloud	−7.481[a]	.000	−5.336[a]	.000	−2.001[b]	.045	−.219[b]	.827
confidence-neighborscount	−3.612[a]	.000	−7.558[a]	.000	−6.746[b]	.000	−8.147[a]	.000
confidence-neighborsrating	−1.798[a]	.072	−2.988[b]	.003	−2.639[a]	.008	−3.067[a]	.002
confidence-perstagcloud	−9.907[a]	.000	−1.069[a]	.285	−4.763[a]	.000	−2.275[b]	.023
confidence-piechart	−6.683[a]	.000	−5.004[b]	.000	−4.113[a]	.000	−.546[a]	.585
confidence-rated4+	−4.765[b]	.000	−9.041[b]	.000	−3.498[a]	.000	−.794[b]	.427
confidence-tagcloud	−7.568[a]	.000	−.656[b]	.512	−1.982[a]	.048	−.435[b]	.663
neighborscount-neighborsrating	−2.445[b]	.014	−8.920[b]	.000	−7.980[a]	.000	−7.663[b]	.000
neighborscount-perstagcloud	−6.174[a]	.000	−4.922[b]	.000	−7.853[a]	.000	−7.666[b]	.000
neighborscount-piechart	−2.733[a]	.006	−9.896[b]	.000	−8.081[a]	.000	−7.558[b]	.000
neighborscount-rated4+	−6.741[b]	.000	−13.333[b]	.000	−8.294[a]	.000	−8.209[b]	.000
neighborscount-tagcloud	−3.515[a]	.000	−6.793[b]	.000	−7.183[a]	.000	−7.346*b*	.000
neighborsrating-perstagcloud	−8.295[a]	.000	−3.520[a]	.000	−2.712[a]	.007	−3.903[b]	.000
neighborsrating-piechart	−5.385[a]	.000	−3.308[b]	.001	−2.454[a]	.014	−2.059[b]	.040
neighborsrating-rated4+	−5.625[b]	.000	−7.487[b]	.000	−1.027[a]	.305	−3.107[b]	.002
neighborsrating-tagcloud	−5.904[a]	.000	−1.871[a]	.061	−.163[b]	.870	−2.485[b]	.013
perstagcloud-piechart	−4.870[b]	.000	−5.423[b]	.000	−1.301[b]	.193	−2.759[a]	.006
perstagcloud-rated4+	−11.895[b]	.000	−8.550[b]	.000	−2.000[b]	.045	−1.377[a]	.169
perstagcloud-tagcloud	−2.831[b]	.005	−2.333[b]	.020	−3.829[b]	.000	−2.323[a]	.020
piechart-rated4+	−10.006[b]	.000	−4.600[b]	.000	−1.028[b]	.304	−1.272[b]	.203
piechart-tagcloud	−2.216[a]	.027	−3.822[a]	.000	−1.968[b]	.049	−1.054[b]	.292
rated4+-tagcloud	−10.134[a]	.000	−8.146[a]	.000	−.981[b]	.327	−.097[b]	.923

Table C.14: Wilcoxon Signed-Rank test. a. Based on neg. ranks. b. Based on pos. ranks.

Appendix D. Interview Plan for the Validation Study

The interviews were conducted in German. The interview plan was structured in four phases and can be summarized as follows. Beside the main questions listed below, additional clarifying questions like "Can you give examples?" were asked when appropriate.

1. The role of explanations. The participants were briefly introduced into the purpose of a recommender system and popular examples for such systems were given. The participants were then informed that some of these systems provide explanatory information when they recommend items. Without giving more details, the main questions in this first phase were:

(1) Please describe the concept "explanations for recommendations" in your own words.
(2) What constitutes a good explanation in your view?
(3) What could be the possible goals of providing explanations?
(4) What is the most important goal in your view?

2. The role of transparency. The term "transparency of recommendations" and the potential effect that users develop more trust in the recommendations when they are accompanied with adequate explanation was explained to the subjects. The following questions were asked.

(1) Is it important to you to understand the reasons why a particular item was recommended?
(2) To which extent would you be willing to invest time in the inspection of explanations in order to understand why something was recommended?

3. Information needs for explanations. The participants were given an example, where a good explanation can help the recipient of a recommendation make a better decision, for instance, when the explanation contains a summary of pro and con arguments. The following question was asked.

(1) What kind of information should an explanation contain to serve as a good basis for decision making?

4. Conclusion of the interview. The following questions were asked.

(1) Can you imagine other possible goals of explanations for recommender systems which we did not discuss so far?
(2) Is there anything else that you want to add on the topic explanations for recommender systems?

References

Bilgic, M., Mooney, R. J., 2005. Explaining recommendations: Satisfaction vs. promotion. In: Proceedings of the Workshop on the Next Stage of Recommender Systems Research (Beyond Personalization'05). San Diego, CA, USA, pp. 13–18.

Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., Riedl, J. T., 2003. Is seeing believing? How recommender system interfaces affect users' opinions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'03). Ft. Lauderdale, Florida, USA, pp. 585–592.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30.

Dias, M. B., Locher, D., Li, M., El-Deredy, W., Lisboa, P. J., 2008. The value of personalised recommender systems to e-business: A case study. In: Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys'08). Lausanne, Switzerland, pp. 291–294.

Felfernig, A., Friedrich, G., Jannach, D., Zanker, M., 2007. An integrated environment for the development of knowledge-based recommender applications. International Journal of Electronic Commerce 11 (2), 11–34.

Friedrich, G., Zanker, M., 2011. A taxonomy for generating explanations in recommender systems. AI Magazine 32 (3), 90–98.

Gedikli, F., Ge, M., Jannach, D., 2011. Understanding recommendations by reading the clouds. In: Proceedings of the 12th International Conference on Electronic Commerce and Web Technologies (EC-Web'11). Toulouse, France, pp. 196–208.

Gedikli, F., Jannach, D., 2010. Rating items by rating tags. In: Proceedings of the 2nd Workshop on Recommender Systems and the Social Web (RSWEB'10). Barcelona, Spain, pp. 25–32.

Gedikli, F., Jannach, D., 2013. Improving recommendation accuracy based on item-specific tag preferences. ACM Transactions on Intelligent Systems and Technology 4.

Herlocker, J. L., Konstan, J. A., Riedl, J. T., 2000. Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW'00). Philadelphia, Pennsylvania, USA, pp. 241–250.

Jannach, D., Hegelich, K., 2009. A case study on the effectiveness of recommendations in the mobile internet. In: Proceedings of the 2009 ACM Conference on Recommender Systems. New York, pp. 205–208.

Jannach, D., Zanker, M., Felfernig, A., Friedrich, G., 2010. Recommender Systems - An Introduction. Cambridge University Press.

Jannach, D., Zanker, M., Fuchs, M., 2009. Constraint-based recommendation in tourism - A multi-perspective case study. Information Technology & Tourism 11 (2), 139–156.

Laforest, J., Bouchard, L., Institut national de santé publique du Québec Staff, 2009. Guide to Organizing Semi-Structured Interviews with Key Informant: Safety Diagnosis Tool Kit for Local Communities. Charting a course to safe living. Gouvernement du Quebec, Ministeres des Communications.

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

Marlin, B. M., Zemel, R. S., Roweis, S., Slaney, M., 2007. Collaborative filtering and the missing at random assumption. In: Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI'07). Vancouver, BC, Canada, pp. 267–275.

McCarthy, K., Reilly, J., McGinty, L., Smyth, B., 2004. Thinking positively - Explanatory feedback for conversational recommender systems. In: Proceedings of the European Conference on Case-Based Reasoning (ECCBR'04). Madrid, Spain, pp. 115–124.

McCarthy, K., Reilly, J., McGinty, L., Smyth, B., 2005. Experiments in dynamic critiquing. In: Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI'05). San Diego, CA, USA, pp. 175–182.

McSherry, D., 2005. Explanation in recommender systems. Artificial Intelligence Review 24 (2), 179–197.

Ong, L. S., Shepherd, B., Tong, L. C., Seow-Choen, F., Ho, Y. H., Tang, C. L., Ho, Y. S., Tan, K., 1997. The colorectal cancer recurrence support (cares) system. Artificial Intelligence in Medicine 11 (3), 175–188.

Porter, M. F., 1997. Readings in information retrieval. Morgan Kaufmann Publishers Inc., Ch. An algorithm for suffix stripping, pp. 313–316.

Pu, P., Chen, L., 2006. Trust building with explanation interfaces. In: Proceedings of the 11th International Conference on Intelligent User Interfaces (IUI'06). Sydney, Australia, pp. 93–100.

Pu, P., Chen, L., August 2007. Trust-inspiring explanation interfaces for recommender systems. Knowledge-Based Systems 20 (6), 542–556.

Rowe, G., Wright, G., April 1993. Expert systems in insurance: A review and analysis. International Journal of Intelligent Systems in Accounting, Finance & Management 2 (2), 129–145.

Sen, S., Harper, F. M., LaPitz, A., Riedl, J. T., 2007. The quest for quality tags. In: Proceedings of the 2007 International ACM Conference on Supporting Group Work (GROUP'07). Sanibel Island, Florida, USA, pp. 361–370.

Sen, S., Vig, J., Riedl, J. T., 2009. Tagommenders: Connecting users to items through tags. In: Proceedings of the 18th International World Wide Web Conference (WWW'09). Madrid, Spain, pp. 671–680.

Senecal, S., Nantel, J., 2004. The influence of online product recommendations on consumers' online choices. Journal of Retailing 80 (2), 159–169.

Swearingen, K., Sinha, R., 2002. Interaction design for recommender systems. In: Proceedings of the 4th Conference on Designing Interactive Systems (DIS'02). London, UK.

Symeonidis, P., Nanopoulos, A., Manolopoulos, Y., 2009. MoviExplain: A recommender system with explanations. In: Proceedings of the 2009 ACM Conference on Recommender Systems (RecSys'09). pp. 317–320.

Thompson, C. A., Göker, M. H., Langley, P., March 2004. A personalized system for conversational recommendations. Journal of Artificial Intelligence Research 21 (1), 393–428.

Tintarev, N., Masthoff, J., 2007a. Effective explanations of recommendations: User-centered design. In: Proceedings of the 2007 ACM Conference on Recommender Systems (RecSys'07). Minneapolis, MN, USA, pp. 153–156.

Tintarev, N., Masthoff, J., 2007b. A survey of explanations in recommender systems. In: Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop (ICDEW'07). Washington, DC, USA, pp. 801–810.

Tintarev, N., Masthoff, J., 2008. The effectiveness of personalized movie explanations: An experiment using commercial meta-data. In: Proceedings of the 5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH'08). Hannover, Germany, pp. 204–213.

Tintarev, N., Masthoff, J., 2011. Designing and evaluating explanations for recommender systems. In: Recommender Systems Handbook. pp. 479–510.

Tintarev, N., Masthoff, J., 2012. Evaluating the effectiveness of explanations for recommender systems - Methodological issues and empirical studies on the impact of personalization. User Modeling and User-Adapted Interaction 22 (4–5), 399–439.

Tuijnman, A. C., Keeves, J. P., 1994. Path analysis and linear structural relations analysis. In: The International Encyclopaedia of Education, 2nd Edition. Oxford, Pergamon, pp. 4229–4252.

Vig, J., Sen, S., Riedl, J. T., 2009. Tagsplanations: Explaining recommendations using tags. In: Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI'09). Sanibel Island, Florida, USA, pp. 47–56.

Vig, J., Soukup, M., Sen, S., Riedl, J. T., 2010. Tag expression: Tagging with feeling. In: Proceedings of the 23nd ACM Symposium on User Interface Software and Technology (UIST'10). New York, NY, USA, pp. 323–332.

Zanker, M., 2012. The influence of knowledgeable explanations on users' perception of a recommender system. In: ACM Conference on Recommender Systems, RecSys '12. Dublin, Ireland, pp. 269–272.

Zanker, M., Bricman, M., Gordea, S., Jannach, D., Jessenitschnig, M., 2006. Persuasive online-selling in quality & taste domains. In: Proceedings of the 7th International Conference on Electronic Commerce and Web Technologies (EC-Web'06). Krakow, Poland, pp. 51–60.

Zhang, J., Pu, P., 2007. A recursive prediction algorithm for collaborative filtering recommender systems. In: Proceedings of the 2007 ACM Conference on Recommender Systems (RecSys'07). Minneapolis, MN, USA, pp. 57–64.