# Understanding Recommendations by Reading the Clouds

Fatih Gedikli, Mouzhi Ge, and Dietmar Jannach

Technische Universität Dortmund,
44221 Dortmund, Germany
firstname.lastname@tu-dortmund.de

**Abstract.** Current research has shown the important role of explanation facilities in recommender systems based on the observation that explanations can significantly influence the user-perceived quality of such a system. In this paper we present and evaluate explanation interfaces in the form of *tag clouds*, which are a frequently used visualization and interaction technique on the Web. We report the result of a user study in which we compare the performance of two new explanation methods based on personalized and non-personalized tag clouds with a previous explanation approach. Overall, the results show that explanations based on tag clouds are not only well-accepted by the users but can also help to improve the efficiency and effectiveness of the explanation process. Furthermore, we provide first insights on the value of personalizing explanations based on the recently-proposed concept of item-specific tag preferences.

**Keywords:** recommender systems, collaborative filtering, explanations, tag clouds, tag preferences.

## 1 Introduction

The capability of a recommender system (RS) to explain the underlying reasons for its proposals to the user has increasingly gained in importance over the last years both in academia and industry. Amazon.com, for example, as one of the world's largest online retailers, allows their online users not only to view the reasons for its recommendations but also to influence the recommendation process and exclude individual past purchases from the recommendation process.

Already early research studies in the area – such as the one by Herlocker et al. [1] – have shown that the provision of explanations and transparency of the recommendation process can help to increase the user's acceptance of collaborative filtering RS. Later on, Tintarev and Masthoff [2] analyzed in greater detail the various goals that one can try to achieve with the help of an explanation facility. Among other aims, good explanations could help the user to make his or her decision more quickly, convince a customer to buy something, or develop trust in the system as a whole.

The question, however, is not only what makes a *good* explanation but also how can we automatically construct explanations which are *understandable* for

C. Huemer and T. Setzer (Eds.): EC-Web 2011, LNBIP 85, pp. 196–208, 2011.

the online user. With respect to the second aspect, Herlocker et al. for example experimented with different visual representations such as histograms of the user's neighbors' ratings. Later, Bilgic and Mooney [3] however observed that such neighborhood-style explanations are good at promoting items but make it harder for users to evaluate the true quality of a recommended item. Thus, they introduced a different, text-based explanation style ("keyword-style explanations") in order to overcome this problem which can in the long term lead to dissatisfaction with the system.

In this work we propose to use *tag clouds* as a means to explain the recommendations made by an RS because tag clouds have become a popular means in recent years to visualize and summarize the main contents, e.g., of a web page or news article. Our hypothesis is that tag clouds are more suitable than keyword-style explanations to achieve the following typical goals of an explanation capability: user satisfaction, efficiency, and effectiveness. As a whole, by achieving these goals, we aim to also increase the users' overall *trust* in the RS.

The paper is organized as follows. In Section 2, we summarize previous works in the area. Section 3 describes the different explanation interfaces, which we evaluated in a user study. Details of the study as well as the discussion of the results are finally presented in Sections 4 and 5 respectively.

## 2 Previous Works

The concept of explanation has been widely discussed in the research of intelligent systems, especially in knowledge-based systems. An explanation facility enables a system to provide understandable decision support and an accountable problem solving strategy. Therefore explanation is considered as one of the important and valuable features of knowledge-based systems [4]. In recent years, the concept of explanations has also been studied and adopted in the area of recommender systems. 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 [1] or enabling more advanced communication patterns between a selling agent and a buying agent [5].

To clarify the goals of providing explanations in recommender systems, Tintarev and Masthoff [2] conduct a systematic review and identify seven goals: transparency (explaining why a particular recommendation is made), scrutability (allowing interaction between user and system), trust (increasing the user's confidence in the system), effectiveness (helping the users make better decisions), persuasiveness (changing the user's buying behavior), efficiency (reducing the time used to complete a task) and satisfaction (increasing usability and enjoyment). In this paper, we propose novel explanation methods and analyze them in line with four of these goals: efficiency, effectiveness, satisfaction and trust.

*Efficiency* means the ability of an explanation to help decreasing the user's decision-making effort. One direct measurement is to compute the time difference of completing the same task with and without an explanation facility or across different explanation facilities. For example, in the user study of Pu and Chen [6], the authors present two different explanation interfaces to users and compared

the time needed to locate a desired item in each interface. *Effectiveness* relates to whether an explanation helps users making high-quality decisions. One possible approach to measure effectiveness is to examine if the user is satisfied with his or her decision. Besides, persuasiveness can be inferred from the study of effectiveness. Vig et al. [7] present four kinds of explanations to users and let users rate how well different explanations help the users decide whether they like a recommended item. An explanation which helps user make better decisions, is considered effective. Compared with persuasiveness, Bilgic and Mooney [3] argue that effectiveness is more important than persuasiveness in the long run as greater effectiveness can help to establish trust and attract users. *Satisfaction* refers to the extent of how useful and enjoyable an explanation helps the users to assess the quality of a recommended item. In the context of recommender system, *trust* can be seen as a user's willingness to believe in the appropriateness of the recommendations and making use of the recommender system's capabilities [8]. Trust can thus be used to determine the extent of how credible and reliable the system is. Tintarev and Masthoff [2] admit the difficulty of measuring trust and suggest measuring it through user loyalty and increased sales. We believe that it is also possible to implicitly examine trust by inferring it from the positive effects of efficiency, effectiveness and satisfaction.

Additionally, note that developing high-quality explanations in recommender systems can further profit from considering different views from related research communities such as intelligent systems, human-computer interaction and information systems. In this paper, we therefore extend the works of [9] and [10] and study the state-of-the-art user interface of tag clouds. Using this interface, we aim to provide an innovative and personalized user interface to achieve higher recommender quality.

## 3 Explanation Interfaces

In this section we will provide an overview of the three different explanation interfaces, which were evaluated in this work: keyword style explanations (KSE), tag clouds (TC), and personalized tag clouds (PTC). KSE, which relies on automatically extracted keywords from item descriptions, is used as the baseline method because this visualization approach has performed best according to effectiveness in previous work. The new methods TC and PTC, however, make use of user-contributed tags, which are a highly popular means of organizing and retrieving content in the Web 2.0.

**Keyword-Style Explanations (KSE).** The KSE interface as shown in Figure 1 has performed the best in the study by Bilgic and Mooney [3]. The interface consists of a top-20 list of keywords, which are assumed to be the most important ones for the user. Note that KSE – in contrast to the other interfaces – does not make use of user-generated tags at all. Instead, it relies on keywords that are automatically extracted from the content description of each item. Internally, an item description has different "slots". Each slot represents a "bag of words",

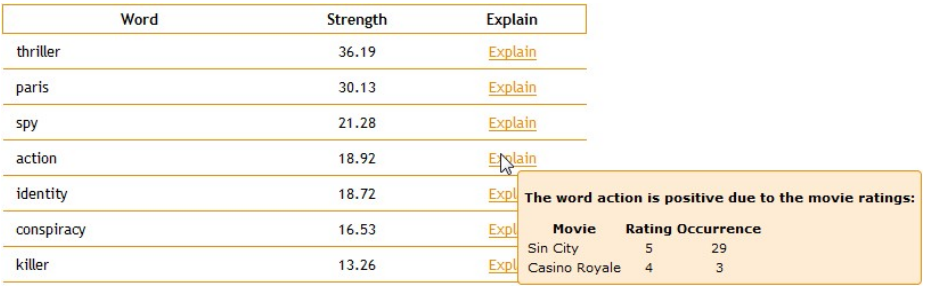



**Fig. 1.** Keyword style explanation (KSE)

that is, an unordered set of words together with their frequencies. Since we are considering the movie domain in our study, we organize a movie's content description using the following five slots: director, actors, genre, description and related-titles. We have collected relevant keywords about director, actors, genre and related-titles from the IMDb website and the MovieLens data set[1]. The data for the description slot was collected by crawling movie reviews in Amazon as well as synopsis information collected from Amazon, Wikipedia and moviepilot[2].

In the KSE-style approach, the importance of a keyword is calculated using the following formula: $strength(k) = t * userStrength(k)$, where $t$ stands for the number of times the keyword $k$ appears in slot $s$. The function $userStrength(k)$ expresses the target user's affinity towards a given keyword. This aspect is estimated by measuring the odd ratios for a given user, that is, how much more likely a keyword will appear in a positively rated example than in a negatively rated one. More formally: $P(k|\text{positive classification})/P(k|\text{negative classification})$. A naïve Bayesian text classifier is used for estimating the probabilities. More details about the KSE-style interface are given in [3].

Beside the list of important keywords, the KSE explanation interface provides a link ("Explain") for each keyword that opens a pop-up window containing all the movies that the user has rated which contain the respective keyword. In this pop-up window the user is provided with both the user's past rating for the movie and the number of times the keyword appears in the corresponding slot.

Note that in [3], the KSE approach performed best in the book domain with respect to effectiveness. However, the evaluation of efficiency and satisfaction was not part of their work but will be analyzed in our study.

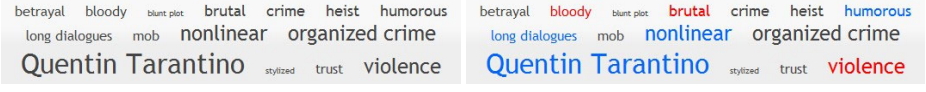
**Tag Clouds (TC).** Tag clouds as shown in Figure 2 (a) have become a frequently used visualization and interaction technique on the Web. They are often incorporated in Social Web platforms such as Delicious and Flickr[3] and are used to visually present a set of words or user-generated tags. In such tag clouds, attributes of tags such as font size, weight or color can be varied to represent

[1] http://www.imdb.com, http://www.grouplens.org/node/73
[2] http://www.amazon.com, http://www.wikipedia.org, http://www.moviepilot.de
[3] http://www.del.icio.us, http://www.flickr.com

relevant properties like relevancy or frequency of a keyword or tag. Additionally, the position of the tags can be varied. Usually, however, the tags in a cloud are sorted alphabetically from the upper left corner to the lower right corner.

In our basic approach of using tag clouds as a not-yet-explored means to explain recommendations, we only varied the font size of the tags, i.e., the larger the font size, the stronger the importance of the tag. We simply used the number of times a tag was attached to a movie as a metric of its importance. The underlying assumption is that a tag which is often used by the community is well-suited to characterize its main aspects. For all other visual attributes we used the standard settings (font sizes etc.). In our future work we also want to analyze the influence of these attributes in explanation scenarios.

Figure 2 (a) shows an example for a movie explanation using the TC interface. Tags such as "Quentin Tarantino" or "violence" have been used by many people and are thus displayed in a larger font size.




(a) Tag cloud (TC) (b) Personalized tag cloud (PTC)

**Fig. 2.** Tag cloud explanation interfaces

**Personalized Tag Clouds (PTC).** Figure 2 (b) finally shows an interface called personalized tag cloud (PTC), which unlike the TC interface is able to exploit the concept of item-specific *tag preferences* [11,12]. The idea of tag preferences is that users should be allowed to assign preferences to tags in order to express their feelings about the recommendable items in more detail. Thus users are not limited to the one single overall vote anymore. In the movie domain, tag preferences can give us valuable information about what users particularly liked/disliked about a certain movie, e.g., the actors or the plot. The PTC interface represents a first attempt to exploit such tag preferences for explanation purposes.

In contrast to the TC interface, we vary the color of the tags according to the user's preference attached to the tag. Blue-colored tags are used to highlight aspects of the movie toward which the user has a positive feeling, whereas tags with a negative connotation are shown in red. Neutral tags, for which no particular preference is known, are shown in black. Again, the font size is used to visualize the importance or quality of a tag. An example of the PTC interface for a crime movie is shown in Figure 2 (b). According to the explanation, the user is assumed to like this movie because of its director *Quentin Tarantino*, whereas *violence* and *brutality* are reasons not to watch this movie.

As explanations are usually presented for items which the user does not know yet, we have to first *predict* the user's feeling about the tags attached to a movie. For this purpose, we analyze the tag preference distribution of the target user's nearest neighbors and decide whether the target user will like, dislike or feel neutral about the item features represented by these tags. In order to predict a

preference for a particular tag, the neighbors preferences for this tag are summed up and normalized to our preference scale for tags. Note that in our study users were able to give preferences to tags on a 5-point scale with half-point increments (0.5 to 5). If the normalized preference lies between $[0.5, 2.0]$ or $[3.5, 5.0]$, we will assume negative or positive connotation respectively; otherwise we will assume that the user feels neutral about the tag.

It is important to know that the interfaces KSE and PTC are personalized, whereas TC represents a non-personalized explanation interface.

## 4 Experimental Setup

We have conducted a between-subjects user study in which each subject was confronted with all explanation interfaces presented above. In this section, we will shortly review the experimental setup which consisted of two phases.

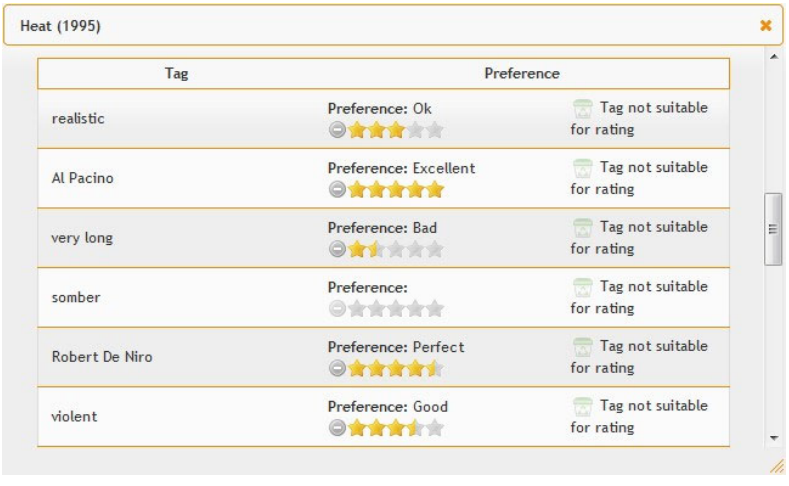



**Fig. 3.** Rating (tags of) the movie *Heat (1995)* on a Likert scale of 0.5 to 5

**Experiment - phase 1.** In the first phase of the experiment, the participants were asked to provide preference information about movies and tags to build the user profiles. The users had to rate at least 15 out of 100 movies[4]. After rating a movie, a screen was shown (Figure 3) in which users could rate up to 15 tags assigned to the movie[5]. On this screen, users could rate an arbitrary number of tags; skip tags, in case they thought that they were not suitable for a given movie; or explicitly mark tags as inappropriate for rating. Note that users were not allowed to apply their own tags as we want to ensure that we have a reasonable overlap in the used tags.

[4] We have limited the number of movies to 100 in order to be able to find nearest neighbors in the PTC approach.

[5] The tags were taken from the "Movie-Lens 10M Ratings, 100k Tags" data set (http://www.grouplens.org/node/73).

**Experiment - phase 2.** In the second phase, which took place a few weeks after the first session, the subjects used an RS[6] which presented them movie recommendations based on the user profile from the first phase. In addition, the different explanation interfaces were shown to the user. In the following, we will introduce our evaluation procedure which extends the procedure proposed by Bilgic and Mooney [3]:

**Procedure 1.** User evaluation

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

The evaluation system randomly selected a tuple $(r, e)$ of possible recommendation and explanation pairs and presented the movie recommendation $r$ using explanation interface $e$ to the end-user without showing the title of the movie. The user was then expected to provide a rating for the movie by solely relying on the information given in the explanation (lines 1-6). The selection order is randomized to minimize the effects of seeing recommendations or interfaces in a special order. If the users recognized a movie based on the information presented in an explanation, they could inform the system about that. No rating for this movie/interface combination was taken into account in this case to avoid biasing effects. We additionally measured the time needed by the users to submit a rating as to measure the *efficiency* of the user interface. These steps were repeated for all movie/interface combinations. Afterwards, we again presented the recommendations to the user, this time showing the complete movie title and links to the corresponding movie information pages at Wikipedia, Amazon and IMDb. We provided information about movies to reduce the time needed for completing the experiment since watching the recommended movies would be too time consuming. The users were instructed to read the detailed information about the recommended movies and then asked to rate the movies again (lines 7-10). According to [3], from the point of view of an end-user, a good explanation system can minimize the difference between ratings provided in the lines 5 (explanation rating) and 9 (actual rating). Thus we can also measure *effectiveness/persuasiveness* by calculating the rating differences. At the end of the experiment, the users were asked to give feedback on the different explanation interfaces (as to measure *satisfaction* with the system) by rating the system as

[6] We used a classical user-based collaborative filtering algorithm.

a whole on a 0.5 (lowest) to 5 (highest) rating scale (line 11). Again, the order was randomized to account for biasing effects.

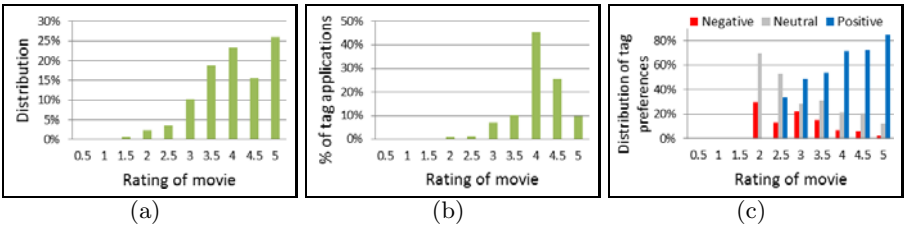
## 5 Empirical Evaluation

### 5.1 Participants

We recruited 19 participants (four female) from five different countries. Most of them were students at our institution with their age ranging from 22 to 37 (average age was 28 years). Ten participants declared high interest in movies, whereas eight were only to a certain extent interested in movies. One person was not interested in movies at all.

### 5.2 Collected Data

The participants provided a total of 353 overall movie ratings and 5,295 tag preferences. On average, each user provided 19 movie ratings and 279 tag preferences and assigned 15 tag preferences to each rated movie. Because participants were also allowed to repeat phase 2 of our user study, we collected a total of 848 explanation ratings (on average 45 ratings per user).




(a) (b) (c)

**Fig. 4.** Distribution of (a) movie ratings, (b) tag applications over movie ratings and (c) negative, neutral and positive tags applied to movies with different ratings

Figure 4 (a) shows the distribution of the movie ratings collected in our study. It can be seen that users preferred to rate movies they liked, i.e., a *positivity bias* is present among the participants which is in line with the findings of other researchers [13,12]. Vig et al. [12] showed that the positivity bias is also present for the taggers, that is, taggers apply more tags to movies they liked compared to movies they rated badly. This finding is also consistent with our results, as shown in Figure 4 (b). Users applied four times more tags to movies they rated with 4 or higher compared to movies to which they gave less than 4 points. Figure 4 (b) shows another interesting effect, which is only partly visible in the data of Vig et al. [12]. Users applied seven times more tags to movies rated with 4 or 4.5 points compared to movies rated with 5 points – the highest rating value – although there are more movies rated with 5 points than with 4 or 4.5

points, as shown in Figure 4 (a). We believe that this effect may be due to users' demand for justifying non-5-point ratings, i.e., users want to explain to the community *why*, in their opinion, a particular movie does not deserve a 5 point rating.

Figure 4 (c) finally shows the distribution of negative, neutral and positive tags applied to movies with different ratings[7]. As expected, a user's movie rating has a strong influence on the tag preferences assigned to a movie. The number of positive (negative) tag preferences increases (decreases) with the overall movie rating. Again, the results are comparable with those reported in [12].

### 5.3 Hypotheses, Results and Discussion

We tested three hypotheses. First, we hypothesized that the tag cloud interfaces TC and PTC enable users to make decisions faster (**H1**:Efficiency). We believe this as we think the visual nature of a tag cloud allows users to grasp the content information inside a cloud faster compared to KSE, which are organized in a more complex tabular structure. We also believe that users enjoy explanations from TC and PTC more than in the KSE style as we assume that tag cloud explanations are easier to interpret for the end user (**H2**:Satisfaction). We further conjecture that users make better buying decisions when their decision is based on TC or PTC rather than KSE (**H3**:Effectiveness). We believe this because we think that compared to TC or PTC, there is a higher risk of misinterpreting KSE because users always have to consider both the keyword and its corresponding numerical importance value, whereas in TC and PTC the importance is encoded in the font size of a tag.

In the following we will have a closer look at the results which are summarized in Table 1. Note that throughout this work we have used the Friedman test with the corresponding post-hoc Nemenyi test as suggested by Demšar [14] for a comparison of more than two systems.

**Table 1.** (a) Mean time for submitting a rating. (b) Mean response of the users to each explanation interface. (c) Mean difference of explanation ratings and actual ratings. Bold figures indicate numbers that are significantly different from the base cases (N is the sample size and $\alpha$ is the significance level).

| | | KSE | TC | PTC | N | $\alpha$ |
|---|---|---|---|---|---|---|
| (a) | Mean time [sec] | 30.72 | **13.53** | **10.66** | 60 | 0.05 |
| | Standard deviation | 19.72 | 8.52 | 5.44 | | |
| (b) | Mean interface rating | 1.87 | **3.74** | **3.87** | 19 | 0.05 |
| | Standard deviation | 0.90 | 0.65 | 0.62 | | |
| (c) | Mean difference | -0.46 | **-0.13** | **-0.08** | 283 | 0.05 |
| | Standard deviation | 1.00 | 1.01 | 1.03 | | |
| | Pearson correlation | 0.54 | 0.79 | 0.83 | | |

[7] For clarity reasons, we have classified the tag preferences into the tag preference groups *negative* ($< 2.5$ points), *neutral* ($2.5 - 3$ points) and *positive* ($> 3$ points).

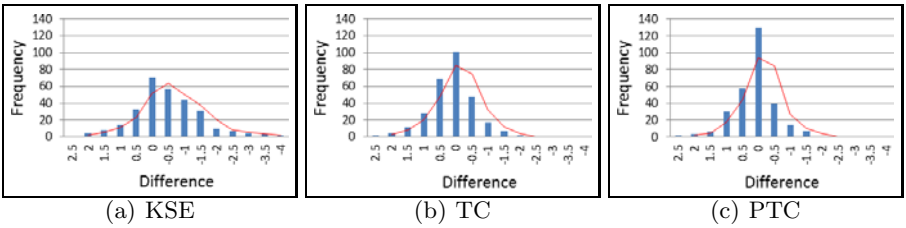
**Efficiency.** To test our hypothesis of improved efficiency of tag clouds, we analyzed the time measurement data which was automatically collected in our study. Table 1 (a) shows the mean times (in seconds) for submitting a rating after seeing the corresponding explanation interface. We can see that the time needed when using the tag cloud approaches is significantly shorter than for KSE. Thus, we can conclude that the data supports hypothesis H1. The data also indicates that the PTC method helps users to make decisions slightly faster than the TC approach, but the difference was not statistically significant.

**Satisfaction.** Table 1 (b) shows the mean response on overall satisfaction of 19 users to each explanation interface based on a Likert scale of 0.5 to 5. It can be seen that users prefer the PTC approach over the TC presentation style and the TC style over the KSE method, which supports hypothesis H2. Again, the differences between the keyword-style explanations and the tag cloud interfaces are significant but no significant difference among the tag cloud interfaces could be found although the data indicates that users favor PTC-style explanations. One possible reason is that tag clouds are in general capable of visualizing the context in a concise manner and can thus help users reduce the time needed to understand the context which in turn increases user satisfaction.

**Effectiveness / Persuasiveness.** Bilgic and Mooney [3] propose to measure effectiveness by calculating the rating differences between explanation rating and actual rating, as described in Section 4. If the difference is 0, the explanation and the actual rating will match perfectly, i.e., the explanation helps the user to accurately predict the quality of an item. Otherwise, if the difference is positive (negative), users will overestimate (underestimate) the quality of an item. In this context we talk about the persuasive power of an explanation system.

Table 1 (c) shows the mean difference of explanation ratings and actual ratings. The histograms showing the mean differences are presented in Figure 5.

The mean differences of the tag cloud interfaces are close to 0 which is an indication that the interfaces are valuable for users to accurately estimate the quality of an item. Note that we have also considered the Pearson correlation between explanation and actual ratings to account for averaging effects. From the user's point of view, a good explanation interface has a mean difference value of 0, a low standard deviation, and a high correlation between both rating values.




(a) KSE (b) TC (c) PTC

**Fig. 5.** Histograms showing the differences between interface and actual ratings

Users can estimate item quality most precisely with the help of the PTC interface. TC explanations are also a good estimator for item quality. The KSE interface has a significantly different value of $-0.46$ which means that KSE cause the user to underestimate the actual rating on average by $-0.46$. On a 5-point scale with half-point increments an underestimation of $-0.46$ on average can be considered as important. Note that in [3], KSE reached a value of 0. We think that the difference in the mean values comes from the different domains considered in our studies (movie domain vs. book domain). Overall the results support our last hypothesis H3.

Next we will discuss about the tradeoff between effectiveness and persuasiveness and the influence of persuasiveness on the user's trust in an RS.

**Trust.** As mentioned above, effectiveness can be measured by the rating difference before and after the consumption or inspection of a recommended item. Smaller differences are indicators of higher effectiveness. Therefore, if the rating for an item based only on the explanation is the same as the rating after the user has consumed the item, we can consider the explanation as highly effective. In the other case, the limited effectiveness will negatively impact on user satisfaction and the trust in the RS.

Consider the following case. A user rates an item with 4 (good) based only on the explanation. After consuming this item, however, the user rates the item with 2 (bad). This means that the user found this item is not as good as expected given only the explanation. In this scenario the user may consider the explanation to be not trustful. We call this effect *positive persuasiveness*, as the system successfully persuades the user to consume/buy the item. Conversely, if the user initially rates the item first with 2 and finally with 4, this means that the explanation does not correctly reflect the truth. In this case, the user may find the explanation to be inaccurate and lose the interest in using this system. We call this effect *negative persuasiveness*. Both positive and negative persuasiveness can cause the loss of trust to users.

The question remains, which form of persuasiveness is better. From a user's perspective, positive persuasiveness may leave the user with the impression that the system is cheating because the system overstates the advantages of the item. This may cause the user to completely abandon the system. However, from a business perspective, if a firm intents to promote a new product or convince the user to adapt a new version of a product, positive persuasiveness may help to increase effects of advertisement and user's familiarity to this product. Negative persuasiveness, on the other hand, has a different effect and may cause the user to suppose that the system does not really take his or her preferences into account. However, we assume it to be a rather "safe" strategy, if we are able to keep the negative persuasiveness level within a certain range. Overall, we argue that it is important to choose the direction of the persuasiveness according to different cases and goals. We can either align positive persuasiveness with the business strategy or control the negative persuasiveness at an acceptable level.

## 6 Summary

In this work, we have presented the results of a user study in which three explanation approaches were evaluated. We have compared keyword-style explanations, which performed best according to effectiveness in previous work, with two new explanation methods based on personalized and non-personalized tag clouds. The personalized tag cloud interface additionally makes use of the recent idea of item-specific tag preferences. We have evaluated the interfaces on the quality dimensions efficiency, satisfaction and effectiveness (persuasiveness) and discussed their impact on the user's trust in an RS.

The results show that users can make better decisions faster when using the tag cloud interfaces rather than the keyword-style explanations. In addition, users generally favored the tag cloud interfaces over keyword-style explanations. This is an interesting observation because users preferred even the non-personalized explanation interface TC over the personalized KSE interface. We assume that there are factors other than personalization such as the graphical representation, which play a crucial role for effective explanation interfaces. The results also indicate that users preferred PTC over TC. We believe that with PTC users need less time to come to an even better conclusion because the font color of a tag already visualizes a user's feeling about the tag and reduces the risk of misinterpreting a tag[8]. Although we view content and the visualization to be tightly interrelated in explanations (as done in previous works), we plan to run experiments in which we evaluate effects of content and visualization separately.

We believe that higher user satisfaction, efficiency, and effectiveness have positive impact on the users' overall trust in the RS which ensures user loyalty and long term wins. In future we want to show in a larger study that the differences between the TC and PTC approaches are significant.

Our future work includes the evaluation of further quality dimensions such as transparency; in addition, we plan to estimate a user's tag ratings automatically in order to reduce the time needed for completing the experiment. This way, we hope to be able to conduct broader studies which involve more test persons.

## References

1. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: CSCW 2000, New York, pp. 241–250 (2000)
2. Tintarev, N., Masthoff, J.: Effective explanations of recommendations: User-centered design. In: RecSys 2007, New York, pp. 153–156 (2007)
3. Bilgic, M., Mooney, R.J.: Explaining recommendations: Satisfaction vs. promotion. In: Beyond Personalization 2005, San Diego, pp. 13–18 (2005)
4. Berry, D.C., Broadbent, D.E.: Explanation and verbalization in a computer-assisted search task. Quart. Journ. of Experim. Psychology 39(4), 585–609 (1987)

[8] For example, consider the case where users see the tags *Bruce Willis* and *romantic movie* in a tag cloud and wonder whether they will like the performance of their action hero in a romantic movie.

5. Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender Systems - An Introduction. Cambridge University Press, Cambridge (2010)
6. Pu, P., Chen, L.: Trust building with expl. interfaces. In: IUI 2006, Sydney, pp. 93–100 (2006)
7. Vig, J., Sen, S., Riedl, J.: Tagsplanations: Explaining recommendations using tags. In: IUI 2009, Sanibel Island, Florida, pp. 47–56 (2009)
8. Cramer, H., Evers, V., Ramlal, S., Someren, M., Rutledge, L., Stash, N., Aroyo, L., Wielinga, B.: The effects of transparency on trust in and acceptance of a content-based art recommender. User Mod. and User-Adap. Inter. 18, 455–496 (2008)
9. Sen, S., Vig, J., Riedl, J.: Tagommenders: Connecting users to items through tags. In: WWW 2009, Madrid, pp. 671–680 (2009)
10. Kim, H.N., Ji, A.T., Ha, I., Jo, G.S.: Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation. Electronic Commerce Research and Applications 9(1), 73–83 (2010)
11. Gedikli, F., Jannach, D.: Rating items by rating tags. In: RSWEB 2010, Barcelona, pp. 25–32 (2010)
12. Vig, J., Soukup, M., Sen, S., Riedl, J.: Tag expression: Tagging with feeling. In: UIST 2010, New York, pp. 323–332 (2010)
13. Marlin, B.M., Zemel, R.S., Roweis, S., Slaney, M.: Collaborative filtering and the missing at random assumption. In: UAI 2007, Vancouver, pp. 267–275 (2007)
14. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)