The effects of transparency on perceived and actual competence of a content-based recommender.

Henriette Cramer[1], Bob Wielinga[1], Satyan Ramlal[1], Vanessa Evers[1],
Lloyd Rutledge[3,4] and Natalia Stash[2]
[1]Human-Computer Studies Lab, University of Amsterdam
Kruislaan 419, 1098 VA, Amsterdam, The Netherlands
hcramer@science.uva.nl
[2]Eindhoven University of Technology, P.O.Box 513, 5600 MD Eindhoven,
The Netherlands
[3] Telematica Institute, Postbus 589, 7500 AN, Enschede, The Netherlands
[4]CWI, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands

ABSTRACT

Perceptions of a system's competence influence acceptance of that system [31]. Ideally, users' perception of competence matches the actual competence of a system. This paper investigates the relation between actual and perceived competence of transparent Semantic Web recommender systems that explain recommendations in terms of shared item concepts. We report an experiment comparing non-transparent and transparent versions of a content-based recommender. Results indicate that in the transparent condition, perceived competence and actual competence (in specific recall) were related, while in the non-transparent condition they were not. Providing insight in what aspects of items triggered their recommendation, by showing the concepts that were the basis for a recommendation, gave users a better assessment of how well the system worked.

Author Keywords

Perceived competence, actual competence, transparency, explanations, recommender systems

ACM Classification Keywords

H5.m. [Information interfaces and presentation (e.g. HCI)]: Miscellaneous. H1.2. Models and Principles: User/Machine Systems.

INTRODUCTION

The amount of information on the World Wide Web is enormous; with increasing chances of information overload and making finding the information you actually need challenging. The Semantic Web is an attempt to overcome these problems. Recommenders are one type of application that can be used to help users find information on the Semantic Web [34, 18]. Recommender systems recommend information items that the specific user would find interesting. Recommendations are based on information gained from implicit or explicit user feedback, typically ratings of other items. Recommendations can be based on how similar items are to items a user previously liked (content-based recommendation), can be based on which items similar users like (collaborative, or social-based recommendation) or can be based on a combination of approaches. Semantic Web techniques such as metadata, annotations and ontologies can be used in recommenders as well [e.g. 5, 9, 23, 35].

One of the challenges in user interaction with Semantic Web applications is making these applications transparent to the user; making 'inner workings', inferences and reasons for results understandable [19]. A content-based recommender system that processes Semantic Web formats can, for example, show which concepts have been used to recommend certain items and which concepts are included in a user's profile. The user can then investigate the competence of a system by comparing the criteria in his or her profile with the criteria he or she actually finds important. Perceived competence of a system has been shown to influence users' acceptance of a system [31]. Lee and See [17] state that appropriate trust and reliance depend on how well the capabilities of an application are conveyed to the user as well. Preferably, perceived competence of a system matches the actual competence of a system. However, this might not always be the case. Making a system understandable to the user might increase the likelihood of appropriate usage decisions. In this study the effects of transparency on the relationship between actual and perceived competence are investigated in the context of a content-based recommender system that uses Semantic Web techniques.

Problem statement

Appropriate usage decisions depend on the user's ability to assess whether using a system is beneficial in his/her situation [17]. This paper investigates whether making a system transparent influences the relation between actual and perceived competence of a Semantic Web recommender system. Focus is on the criteria that underlie recommendations. We expect that when the concepts used to recommend a certain item are transparent to the user, perceived and actual competence are more likely to be related. Transparency is additionally expected to increase competence of the system by helping users to identify and correct system mistakes.

BACKGROUND, RELATED WORK

Semantic Web recommender systems

Semantic Web technologies can be used in recommender systems in various ways. Semantic annotations of information items can, for example, be used for content-based recommendations. Ziegler et al. [35] use taxonomies for classification of products, such as books and movies, for the computation of personalised product recommendations. Celma [5] and Loizou and Dasmahapatra [18] also describe ontology-based recommendations. [3, 5, 9, 23, 35] combine Semantic Web-based techniques with trust networks to generate social-based recommendations based on the trust users will have in other users' recommendations. Even in current, popular recommenders, the role of concepts related to items is gaining increasing importance. Amazon's recommender [1] now uses categories and user tags of its items to facilitate browsing and to fine-tune recommendation lists. MovieLens [22] also lets users tag movies as well as rate tags, although this rating reflects confidence in the annotation rather than the user's interest in its topic. Revyu [25], the winner of the last Semantic Web Challenge [27], lets users make and rate any tag, representing items and topics equivalently, although it performs no recommendation.

Specific domains might offer specific opportunities for using Semantic Web technologies for content-based recommendation. The information needed for such content-based recommendations is widely available in, for example, the cultural heritage domain. Descriptions of artworks and their characteristics and expert's annotations provide for widely available metadata that can be processed to, for example, generate personalised recommendations or for new search strategies [2, 13, 26, 28].

It is important to investigate ways to evaluate whether such systems actually fit user needs and how practical development of satisfying systems can be facilitated.

Acceptance, trust, competence

Perceptions and expectations of a system influence trust and acceptance of a system [24, 31]. According to the Technology Acceptance Model model [31], performance expectancy (including perceived competence), effort expectancy (how much effort it will take to learn and use an application), social influence and facilitating conditions (users' available time and resources) are expected to influence intent to use a system and actual usage behaviour. Trust in a system can be defined as "*the extent to which one party is willing to depend on somebody or something, in a given situation with a feeling of relative security, even though negative consequences are possible*" [Jøsang and Lo Presti, 14]. Trust can be viewed as attitude that both the intentions and goals behind a system and its competence match the user's needs. In order to achieve appropriate trust and reliance on a system, the capabilities of a system need to be conveyed to the user [17]. Focus in this paper is on the perception of competence of a system. When the user cannot assess whether the concepts used by the system match his/her criteria, the interaction will not likely be satisfactory.

Transparency

Transparency entails providing users with ways to increase their understanding of how a system works. Transparency can be offered in various ways and on various levels [11]. Explanations can be offered for why a particular recommendation has been made, why recommendations have been ordered a certain way, or on the general mechanisms that underlie the system. McGuinness et al. [19] list a number of requirements for explanations in a Semantic Web setting. These requirements include interoperability and standardisation of explanation metadata and catering to various types of users and contexts. Their Inference Web approach includes concepts for representing information about trust, information manipulation traces and provenance (data on information sources). Transparency of a system can lead to increased trust and acceptance, an increase in system performance and a more positive user attitude towards using a system [11, 12, 15, 21, 24, 29, 30, 32, 33]. Transparency information can have unexpected side effects as well. Transparency can, for example, affect competence and acceptance of a system negatively if users cannot recognise high-quality profiles [32] or when explanations are poorly designed and difficult to understand themselves [11]. Explanations might additionally make system results more convincing to the user and increase trust in a system, even if the results do not warrant such trust [8]. Ideally, transparency will help appropriate trust and acceptance of a system by increasing user understanding and is not only used to 'promote' system results [4]. Here we investigate whether making a system transparent influences the relation between actual and perceived competence of a system.

METHOD

This study uses data of an experiment that investigated the effects of transparency on user acceptance and trust in a user-adaptive, content-based recommender system [7]. The experiment used a prototype of a content-based art recommender, the CHIP system [www.chip-project.org, 2]. The data is used here of 60 participants who took part in a between-subject experiment. These participants were exposed to one of three different versions of the recommender varying in transparency.

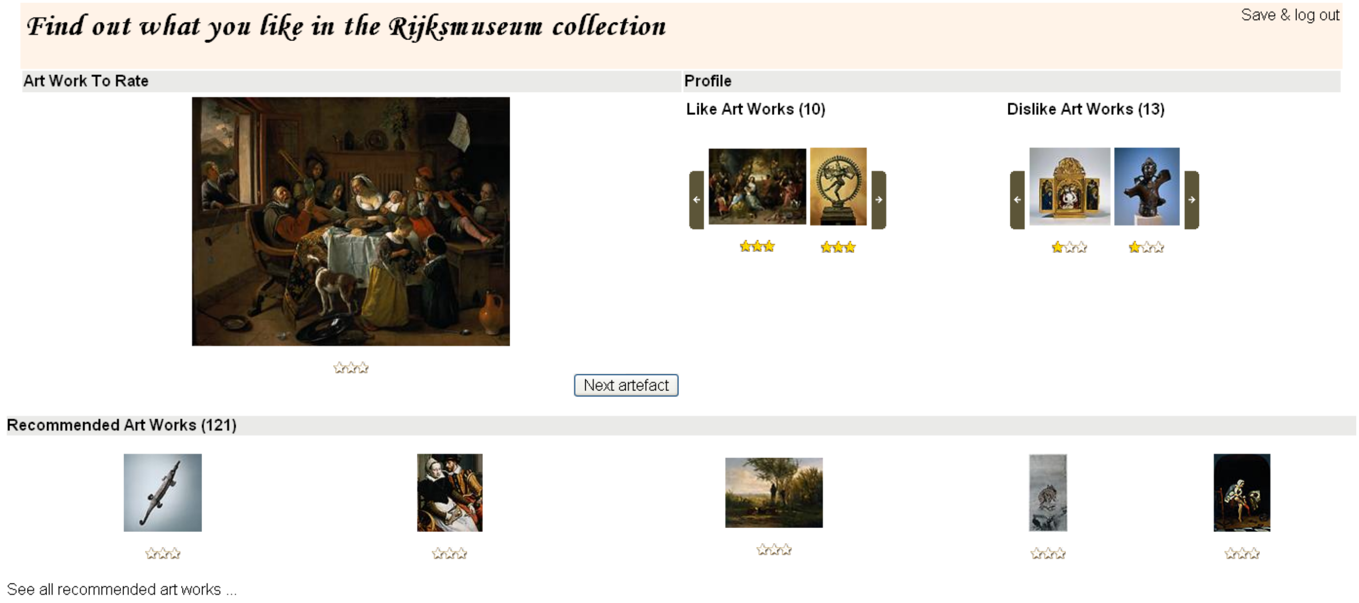

Find out what you like in the Rijksmuseum collection
Save & log out
Art Work To Rate
Profile
Like Art Works (10)
Dislike Art Works (13)
Next artefact
Recommended Art Works (121)
See all recommended art works ...

Fig. 1 Interface non-transparent version

CHIP system

The CHIP system recommends artworks from the Rijksmuseum Amsterdam on the basis of the individual user's ratings of other artworks. It uses content-based, Semantic Web techniques. The artworks have been annotated with a number of concepts such as artist (e.g. Rembrandt), place and time (e.g. Amsterdam, 1700-1750), and themes (such as animals, trompe l'oeil, or musical instruments). These concepts come from standard taxonomy-based vocabularies. Such common vocabularies facilitate fast and consistent annotations and can enable cross-domain recommendation and browsing. The CHIP system offers users the possibility to view and rate Rijksmuseum artefacts, such as paintings and sculptures. The CHIP system calculates predictions of interest in other artworks based on derived user preference patterns for concepts of the rated artworks. In the version used in this study, the user rates individual artworks on how interesting they are. From these ratings, the system generates a user profile using the concepts associated with the artworks. Only these concepts and individual data are used for recommendations. The CHIP project goes beyond database/content-based recommendation, by processing taxonomies in which the metadata sits. This technique involves adding the broader concepts related to the directly annotated item features, similar to the technique proposed by Ziegler [35].

Transparent vs. non-transparent conditions

Each participant interacted with one version of the CHIP system. Three versions of the recommender were used:

- A completely non-transparent version: no information was given on criteria behind the recommendations (Figure 1)
- A 'sure' version. For each recommended artwork, a percentage indicated the confidence of the system that the recommendation is correct (Figure 2). This information was intended to help users realise that a level of uncertainty is involved in recommending. Because participants can compare the confidence rating to their actual interest in the recommendation, this version provided users with information on quality of recommendations. The version was non-transparent on the underlying criteria: no information was given on the reasons for a particular recommendation.
- Transparent 'why' version. The third version was designed to provide insight in the criteria the system used to recommend specific artworks. Each thumbnail of a recommended artwork included a 'why?' link (Figure 3). The 'why' link opened a new window that showed a list of the concepts that the recommended artwork had in common with other artworks that he or she had rated positively (Figure 4).

The three versions only differed in these transparency features in the "recommended Art Works" section of the screen; all other interaction remained the same.

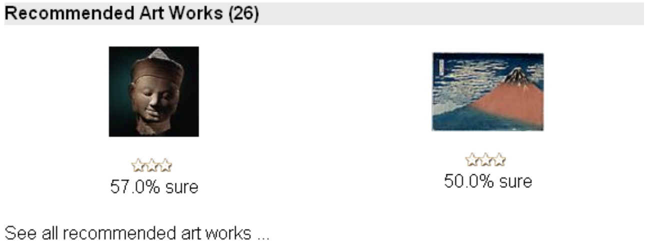


Fig. 2 Detail 'sure' version

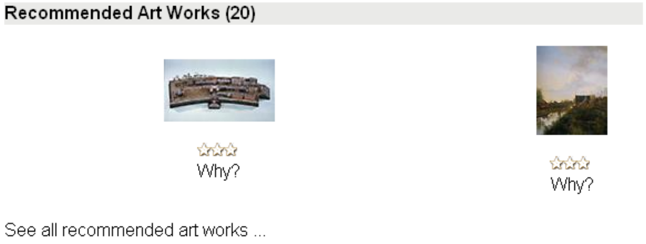
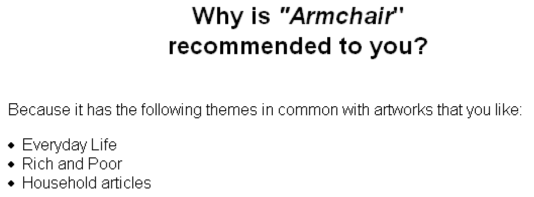


Fig. 3 Detail transparent 'why' version


Why is *"Armchair"* recommended to you?

Because it has the following themes in common with artworks that you like:

- Everyday Life
- Rich and Poor
- Household articles

Fig. 4 Explanation feature 'why' version

Procedure

The data of 60 participants were analysed who participated in one of the three conditions. Participants took part in individual task-oriented sessions, which lasted 45 minutes to three hours. Participants were asked to interact with one of the three versions of the CHIP system. Their ratings of artworks as interesting, neutral or uninteresting were logged. Participants' task was to prepare a (fictional) presentation of their art interests and use the system to find artworks they liked. They were asked to choose 6 artworks they had seen during the interaction with the system's collection of art as their favourites. These artworks were selected from the CHIP system during interaction. Participants then filled out a questionnaire including items on perceived competence of the system. Afterwards, participants were individually interviewed; interviews included a question why they had chosen the 6 artworks as their favourites. Participants' answers were used as representations of their art interests, which were used as input to compute actual competence measures. The study was conducted by two researchers; each individually interviewed 50% of the participants in each of the three conditions. Care was taken to ensure the participants experience with recommender technology, knowledge on art, age, and gender were balanced over the three conditions. In total, actual competence of the recommendation criteria could be determined for 57 participants; 3 participants for whom log files were missing were excluded. 22 participants were included for the non-transparent condition, 18 for the 'sure' condition and 17 for the transparent 'why' condition. Participants in the transparent 'why' condition were explained the transparency feature, but were not actively instructed to use it. In this condition, only participants who did interact with the transparency feature were included.

Table 1 provides an overview of the participant sessions.

Procedure participant session		
1.	Interact with CHIP system	Participants interact with CHIP system and rate artworks. Results in logged user profile consisting of semantic criteria.
2.	Choose Favourites	Participants choose six favourite artworks.
3.	Questionnaire	Participants fill out questionnaire items on perceived competence. Results in perceived competence score.
4.	Interview	Participants provide reasons why they like the chosen six artworks.
5.	Analysis	Compute actual competence by comparing mentioned art interests with semantic concepts in user profile. Results in actual competence score. Compare actual and perceived competence.

Table 1 Procedure participant session

Actual competence measures

Table 2 provides an overview of the measures used in this study. Below we then first discuss the actual competence measures used and the rationale behind them in detail, after which we discuss these details for the perceived competence measures.

Accuracy measures consider whether recommendations are or are not relevant to a user. Both per-person and overall system accuracy measures exist. Accuracy measures empirically measure how close a recommender system's predicted rankings of items for a user differs from the user's true ranking of preference and how well a systems can predict an exact rating for an item [10]. Classification metrics, for example, are a type of accuracy metrics that "measure the frequency with which a recommender system makes correct or incorrect decisions about whether an item is good" [10]. Accuracy measures can also take into account whether recommendations are ranked in the right order.

McNee et al. [20] point out that using only accuracy metrics might not always be suitable to measure competence of a system. They assert that the recommendations that are most accurate according to accuracy metrics might not be the recommendations that are most useful to users. Users might not always have the same needs every time they make use of the same recommender; they might become more experienced and might have different needs depending on their context. McNee et al. note that recommendations should be judged on whether such user needs are met. They also point out that recommendations that might achieve the same accuracy on traditional metrics might be perceived differently by users.

Data, measures	
Perceived competence	
User Interests	Interests named in interview
Questionnaire	8 Likert-type scale questionnaire items e.g. "I think that the system's criteria in choosing recommendations for me are similar to my own criteria."
Actual competence	
Profile interests	Semantic concepts included in user's profile used by the system to recommend artworks (logged)
Number of profile concepts matching user interests	Underlying concepts used by the system to recommend artworks match the interests mentioned by participant
Precision	Number of profile concepts matching with user interests / number of concepts in profile
Recall	Number of profile concepts matching with user interests / number of interests mentioned by participant
F-score	2 x (precision x recall) / (precision + recall).

Table 2 Overview measures and data sources

Accuracy measures used in this study

Most accuracy metrics focus on the fit of recommendation results. Herlocker et al.'s [10] discussion for example also focuses on whether a user would rate a recommended item as interesting or uninteresting. In contrast we here focus on the semantic-derived criteria that underlie content-based recommendation results. Indeed it is very important that resulting recommendations are appropriate, but this is not likely to be the case if the underlying criteria do not fit the user's needs. Of interest here is how providing users insight in these concepts might change their perceptions of competence of the system. Therefore we measure both actual (underlying) competence and perceived competence.

A specific measure for actual competence has been used in this study. The measure used here investigates whether the underlying concepts used by the system to recommend artworks match the participant's interests. In other words, it investigates whether the concepts the system thinks the participant finds interesting, actually do interest the participant. For the purposes of this study, per-person measures were needed to be able to correlate actual and perceived competence scores.

During the interview, each participant was asked for each of the six artworks he or she had chosen why this artwork was interesting to them. Their answers were taken as a representation of their interests in art. Example statements included 'my dad had a collection of these", "dark colours", "history behind {the artwork}", "Amsterdam". For each participant, the list of interests was divided into unique statements on concepts of artworks he or she liked. Duplicate interests were removed. Statements such as "retains my attention", "cute", "interesting", "would hang it on my wall", were removed. These statements did convey that the user liked the artwork or found it interesting, but did not provide information on why this was the case. However, exactly these reasons why were of interest here, hence the removal of these statements.

Duplicate concepts or concepts that appeared subclasses of other concepts were removed. For example, a property like "Homes, 17th-century" was taken as a subclass of "domestic interiors" and not taken included in the final score. If, for example, the system showed "Gods and deities" as a property, all concepts referring to deities, such as "Zeus" or "Poseidon" were removed as well. Matching was then applied flexibly. A participant's mentioning of liking "Nice little still lifes of vegetables and fruit" would be marked as a match if the system showed a property like "food and drink", or "still life" and not necessarily both of them. "Hinduism" was interpreted as a positive match with "Asian art" if one of the artworks the participants liked was, for example, a Hinduism-related statue.

Also removed were statements that were hard to interpret for the researchers e.g. "aesthetics, can't really explain it", as it was hard to identify what type of concept would fit these statements. Statements such as "nicely made", "lots to see" were used if they were interpretable in the context of the participant's other statements to match, for example, concepts such as "painting techniques" or "conversation pieces". For some of the statements made by participants it was not completely clear whether they wanted to explain something about the painting or about their interests. When someone would pick the Nightwatch as a favourite and they would mention "a Rembrandt" in answering the question why they liked the painting, Rembrandt was interpreted as one of the interests.

To investigate the competence of the recommendation criteria, the user profile the system had built up was compared to reported art interests. The number of matches between the unique concepts in a user's profile and his or her actually mentioned interests were used to calculate per-person measures of precision and recall. To sustain reliability of the assessment data, two coders both individually rated for each participant whether the unique concepts in the user profile and the interests a participant had mentioned matched.

Metrics

As measures for actual competence in this study, measures were adapted from the traditional information retrieval competence measures of precision, recall and f-score. These measures are based on the proportion of relevant documents a system presents the user with. Items or concepts are rated as either relevant or non-relevant.

Precision of the system's user profile was computed by:

number of profile concepts matching with user interests / number of concepts in profile

A perfect score (1) for precision means that no irrelevant concepts have been included in the user's profile. Precision scores will be quite low in this study, as it is quite possible not all art interests of the participant are captured in the six artworks they chose as their favourites. These scores should not be taken as absolutes, but will serve their purposes as comparative measures for this study.

Recall was computed with the following formula:

number of profile concepts matching with user interests / number of interests named by participant

If recall has a value of 1, all of a participant's interests have been included in his or her profile. A value of 0 means that none of the interests mentioned by a participant have been included in his or her profile. Reasons for low recall scores could include: the system's algorithm perhaps not being up to the task, the collection of terms (used to annotate artworks e.g. the ontology) not being up to par, or incomplete or incorrect annotations of artworks.

The f-score measure, a weighed combination of precision and recall, was then used as a score for the "actual competence" construct:

F-score = 2 x (precision x recall) / (precision + recall).

An f-score of 1 would be a perfect score; an f-score of 0 would be the worst score possible.

These measures are used while taking into account the issues raised by McNee et al. [20] and Herlocker et al. [10] in regards to evaluation measures. Herlocker et al. for example, note that recall is impractical to measure in recommender systems. The value of the recall metric depends greatly on how many items have been rated by the user. The value for recall will, for example, be rather low in this study. Pure recall requires knowing whether each item it relevant. This means that all items in the system should be rated by the user. In this case where we look at underlying concepts in the user profile, all possible concepts should have been rated. This is not the case here. Additionally, traditional recall measures assume all relevant items and concepts are present in a system's database. Here artworks that are of interest to the user might not be available in the system and certain art concepts might not have been taken into account.

For the purposes of the analysis here, participants were not asked to rate the recommendations but to provide insight in the underlying criteria they used to choose their favourite artworks. Ideally, to get accurate scores on precision, recall and f-score measures all participants should have been asked to rate all concepts as interesting or uninteresting. In this study, insight into actual user interests was acquired by asking users why they like the six artworks they found most interesting. As participants were only asked to mention why they thought these six favourite artworks were interesting, these lists of mentioned interests will not be a complete overview of participants' interests. In particular the precision score will probably be lower than actual competence. This is not a problem for the purposes of this study, as scores will be lower for all participants. The metrics used here should not be used taken as absolute measures of accuracy of the CHIP system, but rather as comparative measures serving the purposes of this study. Overall correlation between these scores and participants' perceptions of the system are of interest, and not an absolute match. We were most interested in letting participants freely express their interests and the comparison of these interests with profile interests. We were less interested in establishing absolute, ideal metrics.

Perceived competence measures

To be able to relate the actual competence of the system with the perceptions of the participants, a number of Likert-scale questions were included in the questionnaire participants filled out. How competent participants perceived the system to be, was calculated by averaging participant scores on eight questionnaire items such as "I think that the system's criteria in choosing recommendations for me are similar to my own criteria", "The system correctly adapts its recommendations on the basis of my ratings", "I think that the artworks that the system recommends correspond to my art interests" and "I think the system should use other criteria for recommending artworks to me than it uses now" (question inverted for analysis). All items were 7-point scales, ranging from 1 ('very strongly disagree') to 7 ('very strongly agree'). Cronbach's Alpha for the scale was .914, Mean score=4.07, st.dev=1.15, range:1.63-6.50. More information on measures on system perceptions and participant attitude towards the system can be found in [7].

RESULTS AND DISCUSSION

After reporting the interrater reliability for our competence data, we first discuss the actual competence scores of the recommender and the effects of transparency on these scores. We then investigate effects of transparency on the correlation between actual competence and perceived competence of the system.

Interrater reliability for actual competence

Interrater reliability between the two coders was calculated to assess the reliability of the coding and rating process of the interview comments related to participants' art interests and their user profiles. This study reports the intraclass correlation coefficient (ICC) as the interrater reliability measure. The intraclass correlation coefficient can range between 0, indicating no interrater agreement, and 1, indicating total interrater agreement. The intraclass correlation coefficient (two-way mixed, single measure, absolute agreement) for the number of unique matches between the user profile and the interests mentioned by the participant was .873. Having obtained an acceptable interrater reliability, the subsequent analyses were carried out on the data of one coder. A similar procedure is followed by Kurasaki [16].

The effect of transparency on actual competence

Overall, precision, recall and f-scores (table 3) were rather low. This was partly expected due to our method to calculate these scores (see above section 'actual competence measures' for a discussion), as well as due to possible competence problems of the recommender system used.

	N	Min	Max	Mean	Mdn	St.dev
Precision	57	0	1.00	0.177	0.125	.199
Recall	57	0	1.00	0.272	0.250	.232
F-score	57	0	.467	0.158	0.134	.129

Table 3 Average actual competence recommendation criteria: precision, recall and f-scores.

Often user profiles contained many more concepts than participants named, ranging from 0 to 106 unique concepts, with an average of 21. Participants named between 3 and 17 unique interests, with an average of 8.3. This was not surprising, as participants were not asked to name all of their interests; they were only asked for their particular interest for their 6 favourite artworks from the collection shown to them. The interests named by participants did include interests that were not included in the system concepts, such as use of specific types of colours, or personal collections or hobbies. Length of the user's profile had a significant relation with the recall score (Spearman rho=.645, p(1-tailed)=.000, N=57). This is not surprising; the more terms in a user's profile, the greater the chance the user's interests are captured. Length of the user profile was not related to precision for participants in all conditions.

A Kruskal-Wallis test on differences between the conditions on precision, recall and f-scores revealed a significant difference on precision (H(2)=8.059, p(2-tailed)=.018) and no significant differences on recall (H(2)=1.635, p(2-tailed)=.442) and the derived f-score (H(2)=3.973, p(2-tailed)=.137). Non-parametric procedures are used to test our hypotheses as the data did not meet parametric assumptions. A Mann-Whitney test was used to follow up on the significant finding for precision. It appeared that scores for precision were higher in the non-transparent condition (Mdn=.182) than in the transparent condition (Mdn=.0667), U=105, p(2-tailed)=.020. Precision scores were also higher in the certainty rating 'Sure' condition (Mdn=.174) than in the transparent condition (Mdn=.0667), U=73.5, p(2-tailed)=.007. Recall scores in contrast, were higher in the transparent version (Mdn=.333) than in the 'sure' (Mdn=.236) and non-transparent (Mdn=.174) conditions, but this difference was not significant.

These findings were unexpected. Ideally, transparency could help increase competence by helping users to identify and correct system mistakes; both scores for precision and recall would then be higher by making a system's criteria more insightful to the user. However, this was not the case here. Only differences on precision were significant, and instead of being higher, precision scores in the transparent 'why' condition were lower than in the non-transparent condition. It appears that transparency does not always lead to improved competence. Other possible adverse effects of transparency have been noted by e.g. Waern [32] and Cheverst et al., [6]. It might be that in this case, participants in the 'why' condition noticed that some of their interests were not included in explanations for recommendations and that the transparency feature changed their behaviour. For the purposes of this study, participants could not directly rate concepts; they could only rate artworks. They might have tried to rate artworks that they thought had these missing interesting concepts more positively. It might have been more important for participants in the transparent 'why' condition to try and see all artworks that would be interesting for them, than to eliminate concepts from their profile that they thought were uninteresting. This would be in line with the worry expressed by some participants that when they rated an artwork as uninteresting, the system eliminated too many related artworks that might have been interesting for other, unrelated reasons. However, no significant increase in recall was found for the transparent condition – this is either not an explanation for our finding, or participants were not successful in increasing recall. Direct feedback on concepts in the user profile, while taking into account profiles then have to be truly understandable themselves [32] might be an important feature to ensure positive results when offering transparency of underlying concepts.

The effects of transparency on perceived competence

A Kruskal-Wallis test on differences between the conditions on perceived competence did not yield any significant differences between the conditions (H(2)=2.309, p(2-tailed)=.315). Thus, transparency does not necessary influence the perceptions of competence of a system per se. Further discussion of participant perceptions and their trust in and acceptance of the system in the three conditions can be found in [7].

The effect of transparency on correlations between perceived and actual competence

Spearman rhos were calculated to investigate whether perceived competence was related to actual competence of the recommendation criteria. A comparison of correlations in the different conditions can be found in table 4. For each of the conditions the correlations between perceived competence and the actual competence metrics are given.

		Perceived competence 'why' condition	**Perceived competence** non-transparent condition	**Perceived competence** 'sure' condition
Precision	Rho	.335	.149	.355
	P	.094	.254	.074
	N	17	22	18
Recall	Rho	.519*	.043	.273
	P	.016	.424	.137
	N	17	22	18
F-score	Rho	.390	-.090	.380
	P	.061	.345	.060
	N	17	22	18

Table 4 Correlations actual competence and perceived competence for the three conditions, Spearman rho (1-tailed), * denotes significance.

In none of the conditions a significant correlation was found between perceived competence and precision. However, a significant correlation between perceived competence and recall existed only for participants in the transparent 'why' condition. Making a system more transparent appears to increase the likelihood that users' perceptions of a system's competence match the system's actual competence, at least partially in this case. This appears desirable; we do not want users to perceive a system to be competent when it's not, we want users' perceptions to be representative of its actual competence. It appears that in this case recall played the most prominent role in shaping participants' perception of competence of the system. Perhaps our participants did not change their behaviour following the transparency feature, but the feature appeared useful in assessing system performance. Indeed, participants could not directly correct system mistakes via the transparency feature. This would explain the lack of positive effects on performance, while an effect on the correlation between perceived and actual competence was present.

Table 5 summarises the discussed results.

Effect transparency on:	
Actual competence	
Precision	Highest in non-transparent condition, significant difference. H(2)=8.059, p(2-tailed)=.018. Non-transparent Mdn=.182, 'Sure' Mdn=.174, Transparent 'Why' Mdn=.0667.
Recall	Highest in transparent version, non-significant difference. H(2)=1.635, p(2-tailed)=.442 Non-transparent Mdn=.174, 'Sure' Mdn=.236, Transparent 'Why' Mdn=.333.
Perceived competence	
Perceived competence scores	No significant difference. H(2)=2.309, p(2-tailed)=.315.
Correlations between perceived and actual competence	
Precision	No significant correlation between perceived competence and precision in any of the conditions.
Recall	Significant correlation between perceived competence and recall in transparent 'why' condition only.

Table 5 Overview results

Making underlying concepts understandable to users

It has to be pointed out that we investigated competence of the underlying criteria by interpreting users' interests and trying to match them to expert annotation concepts. A direct matching of system and user concepts was not possible; participants for example used different wording for concepts. Participants in this study indicated they did not always understand all terms given by the system; specialist terms such as 'trompe l'oeil' caused some confusion for some of them. This illustrates that criteria should be meaningful to the user; they should fit both underlying criteria that fit the users' needs and need to be made understandable to the users. Expert annotations might not use the same wording as laymen would, or might use a different level of detail. Different types of users might require different recommendation criteria, but also different types of explanations, geared to their expertise and needs. In a Semantic Web context this might, for example, entail using and matching both laymen and expert ontologies. Additionally, the reasoning processes used in Semantic Web applications can be a lot more complex than the reasoning explained to participants in this study [19]. Issues surrounding understandability of explanations might be even more important when a system makes more complex inferences transparent to the user. Reasoning traces, for example, might not in all cases be understandable to every user. Especially when users are enabled to provide direct feedback on whether the system's reasoning is competent, it is important to make sure the system provides understandable explanations.

CONCLUSION

This paper investigated the relation between actual and perceived competence of a recommender system in transparent and non-transparent conditions. The findings in this study indicate that perception of system competence can differ from actual competence measures. To achieve user satisfaction it is important to focus on user perceptions of the system, and how to match these perceptions with the actual competence of the system.

Concerning the perceived competence of the system, participants appeared to base their perception of competence on recall rather than precision. We found that transparency does not necessarily increase competence. It is important to further explore how explanations can be combined with more direct user feedback to increase competence, especially when more complex reasoning is used to reach results.

Making a system more transparent does increase chances that user perceptions of competence and actual competence of a system are related. In the transparent condition, perceived competence was related to recall (as one measure of actual competence), while in the non-transparent and 'sure' condition it was not related to actual competence at all. We have shown here that transparency indeed is important to include in a system to allow users to calibrate their perceptions to the actual capabilities of a system. This supports the importance of Semantic Web explanation efforts, e.g. McGuiness [19] and efforts to make more complex, distributed reasoning understandable to users.

ACKNOWLEDGMENTS

This research is funded by the Interactive Collaborative Information Systems (ICIS) project, nr: BSIK03024, by the Dutch Ministry of Economical Affairs under contract to the Human-Computer Studies Laboratory of the University of Amsterdam. The CHIP (Cultural Heritage Information Personalization - www.chip-project.org) project is part of the CATCH program funded by the NWO (Netherlands Organisation for Scientific Research).

REFERENCES

1. Amazon, www.amazon.com/gp/yourstore/
2. Aroyo, L., Stash, N., Wang, Y., Gorgels, P. and Rutledge, L. CHIP Demonstrator: Semantics-driven Recommendations and Museum Tour Generation. In Proc. ISWC 2007, Springer (2007), 879-886.
3. Bedi, P., Kaur, H. and Marwaha, S. Trust Based Recommender System for Semantic Web. In Proc. IJCAI 2007, (2007), 2677-2682.
4. Bilgic, M. and Mooney, R.J. Explaining Recommendations: Satisfaction vs. Promotion, In Proc. Beyond Personalization Workshop at IUI 2005, (2005).
5. Celma, O. Foafing the Music: Bridging the semantic gap in music recommendation, In Proc. ISCW 2006, (2006).
6. Cheverst, K., Byun, H.E., Fitton, D., Sas, C., Kray, C. and Villar, N. Exploring Issues of User Model Transparency and Proactive Behaviour in an Office Environment Control System. User modeling and User-adapted interaction 15, 3-4 (2005), 235–273.
7. Cramer, H., Ramlal, S., Evers, V., Van Someren, M. Wielinga, B., Rutledge, L., Aroyo, L. and Stash, N. User interaction with a content-based art recommender. The effects of transparency on trust and acceptance, accepted for publication.
8. Dzindolet, M. The role of trust in automation reliance, Int. J. of Human-Computer Studies 58, 6 (2003), 697 - 718.
9. Golbeck, J. Semantic Web Interaction through Trust Network Recommender Systems. In Proc. ISWC 2005.
10. Herlocker, J.L., Konstan, J.A., Terveen, L.G., and Riedl, J.T. Evaluating Collaborative Filtering Recommender Systems. ACM TOIS 22, 1 (2004), 5-53.
11. Herlocker, J.L., Konstan, J.A. and Riedl, J.T. Explaining collaborative filtering recommendations, In Proc CSCW 2000, ACM Press (2000), 241-250.
12. Höök, K., Karlgren, J., Waern, A., Dahlbeck, N., Jansson, C. G., Karlgren, K., and Lemaire, B. A Glass Box Approach to Adaptive Hypermedia., User Modeling and User-Adapted Interaction 6, 2–3 (1996), 157–184.
13. Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., and Kettula, S. MuseumFinland - Finnish Museums on the Semantic Web. J. of Web Semantics 3, 2 (2005), 224-241.
14. Jøsang, A. and Lo Presti, S. Analysing the Relationship between Risk and Trust. International Conference on Trust Management, Springer (2004), 135-145.
15. Kay, J. Scrutable adaptation: because we can and must. In Proc. AH 2006, Springer (2006), 11-19.
16. Kurasaki, K. S. Intercoder Reliability for Validating Conclusions Drawn from Open-Ended Interview Data. Field Methods 12, 3 (2000), 179-194.
17. Lee J.D. and See K.A. Trust in automation: designing for Appropriate Reliance. Human Factors 42, 1 (2004), 50–80.
18. Loizou, A. and Dasmahapatra, S. Recommender Systems for the Semantic Web. In Proc. ECAI 2006 Recommender Systems Workshop (2006).
19. McGuinness, D.L., Ding, L., Glass, A., Chang, C., Zeng, H. and Furtado, V. Explanation Interfaces for the

Semantic Web: Issues and Models, In Proc. SWUI 2006, (2006).

20. McNee, S.M., Riedl, J. and Konstan, J. Accurate is not always good: How Accuracy Metrics have hurt Recommender Systems. Ext. Abstracts CHI 2006, ACM Press (2006), 1-5.

21. McSherry, D. Explanation in Recommender Systems. Artificial Intelligence Review 24, 2 (2005), 179–197.

22. Movielens, www.movielens.org

23. Nejdl, W., Ghita, S., Paiu R. Semantically Rich Recommendations in Social Networks for Sharing, Exchanging and Ranking Semantic Context. In Proc. ISWC 2005, (2005).

24. Pu, P. and Chen, L. Trust-Inspiring Explanation Interfaces for Recommender Systems. Knowledge-Based Systems Journal 20, 6 (2007), 542-556.

25. Revyu, www.revyu.com

26. Schreiber, G., Amin, A., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Hollink, L.,Huang, Z., van Kersen, J., de Niet, M., Omelayenko, B., van Ossenbruggen, J., Siebes, R., Taekema, J., Wielemaker, J., and Wielinga, B. J. Multimedian e-culture demonstrator. In Proc. International Semantic Web Conference, (2006), 951-958.

27. Semantic Web Challenge 2007, challenge.semanticweb.org.

28. Sinclair, P., Lewis, P., Martinez, K., Addis, M., and Prideaux, D. Semantic web integration of 33 cultural heritage sources. In Proc. World WideWeb, (2006), 1047-1048.

29. Sinha, R., Swearingen, K. The Role of Transparency in Recommender Systems. In Ext. Abstracts CHI 2002, ACM Press (2002) 830-831.

30. Tintarev, N. and Masthoff, J. Effective explanations of recommendations: user-centered design, In Proc. RecSys 2007, ACM Press (2007), 153-156.

31. Venkatesh, V., Morris, M.G., Davis, G.B., and Davis, F.D. User Acceptance of Information Technology: Toward a Unified View, MIS Quarterly 27, 3 (2003), 425-478.

32. Waern, A. User Involvement in Automatic Filtering: An Experimental Study. User Modeling and User-Adapted Interaction 14, 2-3 (2004), 201–237.

33. Wang, W. and Benbasat, I. Recommendation Agents for Electronic Commerce: Effects of Explanation Facilities on Trusting Beliefs. Journal of Management Information Systems 23, 4 (2007), 217-246.

34. Ziegler, C-N. Semantic Web Recommender Systems, In Selected and Revised Papers EDBT 2004, (2004), 78-89.

35. Ziegler, C-N., Lausen, G. and Schmidt-Thieme, L. Taxonomy-driven Computation of Product Recommendations, In Proc. CIKM 2004, ACM Press (2004), 406 - 415.