Explaining recommendations

Nava Tintarev

A dissertation submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy
of the
University of Aberdeen.

Department of Computing Science

2009

Declaration

No portion of the work contained in this document has been submitted in support of an application for a degree or qualification of this or any other university or other institution of learning. All verbatim extracts have been distinguished by quotation marks, and all sources of information have been specifically acknowledged.

Signed:

Date:

Abstract

Recommender systems such as Amazon, offer users recommendations, or suggestions of items to try or buy. These recommendations can then be explained to the user, e.g. *"You might (not) like this item because..."*. We propose a novel classification of reasons for including explanations in recommender systems. Our focus is on the aim of effectiveness, or decision support, and we contrast it with other metrics such as satisfaction and persuasion. Effective explanations should be helpful in the sense that they help users find items that they like (even after trying them), and discard items they would not like.

In user studies, we found that people varied in the features they found important, and composed a short list of features in two domains (movies and cameras). We then built a natural language explanation generation testbed system, considering these features as well as the limitations of using commercial data. This testbed was used in a series of experiments to test whether personalization of explanations affects effectiveness, persuasion and satisfaction. We chose a simple form of personalization which considers likely constraints of a recommender system (e.g. limited meta-data related to the user) as well as brevity (assuming users want to browse items relatively quickly). In these experiments we found that:

1. Explanations help participants to make decisions compared to recommendations without explanations, we we saw as a significant decrease in opt-outs in item ratings - participants were more likely to be able to give an initial rating for an item if they were given an explanation, and the likelihood of receiving a rating increased for feature-based explanations compared to a baseline.

2. Contrary to our initial hypothesis, our method of personalization could damage effectiveness for both movies and cameras which are domains that differ with regard to two dimensions which we found affected perceived effectiveness: cost (low vs. high), and valuation type (subjective vs. objective).

3. Participants were more satisfied with feature-based than baseline explanations. If the personalization is perceived as relevant to them, then personalized feature-based explanations were preferred over non-personalized.

4. Satisfaction with explanations was also reflected in the proportion of opt-outs. The

opt-out rate for the explanations was highest in the baseline for all experiments. This was the case despite the different types of explanation baselines used in the two domains.

Keywords: *Explanations, recommender systems, human-computer interaction, user-centered design, natural language generation, Amazon.*

Acknowledgements

A thesis is never the work of own person. Firstly, a great big thank you to my two exceptional supervisors: Judith and Ehud. For everything you have taught me, for the joint work, and a great amount of support, patience, and faith. There is not a doubt in my mind that this would have been impossible without you. I am fortunate to know you as colleagues and as friends. Thank you also to the exceptional teachers and tutors who have inspired me along the way. In particular, Bengt Persson and Anna Woollams who fed an ever-growing thirst for knowledge, and who never slowed or dumbed anything down.

Thanks are also due to all the reviewers and feedback given at conferences and elsewhere, without which I would not have been able to grow as much as I did. A special thank you to Jim Burnette, Judy Kay and David Lifson, for dialogues that directly impacted the directions my thesis took.

My eternal gratitude to friends across the globe and locally who put up with frequent absences and long hours. Thank you for your support, but also for reminding me that there are other things going on out there in the big world, especially those of you who kept me supplied with travelogues. Special mention here also to my fellow dancers and my musician friends, and those who distracted me with coffees and hot chocolates.

And of course, thank you to my parents, for teaching me the value of being persistent and working hard, but who also continuously and lovingly fueled an ever growing curiosity be it for books, dance or mathematics. I thank you all for your support, I hope I have made you proud.

Contents

List of Tables

List of Figures

Chapter 1

Overview

Recommender systems such as Amazon, offer users recommendations, or suggestions of items to try or buy. These recommendations can then be explained to the user, e.g. *"You might (not) like this item because..."*.

We begin this thesis by giving reasons for including explanations in recommender systems. Our focus is on the aim of effectiveness, or decision support, and we contrast it with other metrics such as satisfaction and persuasion. User acceptance of recommendations (persuasion) is different from acceptance of a system (satisfaction), both of which are different from acceptance of the actual items (effectiveness).

Effective explanations should be helpful in the sense that they help users to find items that they like (even after trying them), and discard items they would not like. Thus, we investigated the properties of helpful explanations, both in terms of content and presentation, in a number of user studies. In the case of presentation, we only conducted pilot studies of *perceived* helpfulness. The core of the thesis however regards explanation content, and personalization in particular.

Using the findings from our user studies, we built a natural language explanation generation testbed system. This testbed was subsequently used in a series of experiments to test whether personalization of explanations affects effectiveness, persuasion and satisfaction. The experiments were conducted in two domains: movies and cameras. The domain choices were partly motivated by a precedent in previous research, and the availability of relevant data. Also, these two domains are each others polars with regard to two dimensions which affect perceived effectiveness: cost (low vs. high), and valuation type (subjective vs. objective).

This introductory chapter aims to serve as a reading guide for the remainder of the thesis, describing each chapter individually. In addition to this overview, the reader is invited to refer to a more detailed summary at the end of each chapter. A list of related publications can be found in Section 1.4.

1.1 Research questions

- *Why explain?* Explanations have been a current topic in the area of recommender systems, but what *is* the added benefit of explanations? In particular, do explanations help users make decisions, and are these decisions more correct than decisions made without explanations (effectiveness)?

- *Does the type of explanations that we offer to users matter?* More specifically, does personalization of explanations lead to better effectiveness? Are users more satisfied with personalized explanations?

- *How do we measure that explanations were effective?* In particular we investigate the utility and limitations of an existing metric.

1.2 Key contributions

- **Why explain?** As part of the evaluation of explanations, we had to consider what constitutes a good explanation. This resulted in a formulation of a classification for including explanations in a recommender system, defined as aims and metrics, in the context of a thorough literature review (see Chapter 3). While the role of explanations in expert and recommender systems has long been recognized, our literature review identifies reasons for explaining, and recognizes that these are distinct and may even conflict. This is the *first* such analysis of the aims of explanations for recommender systems, and answers the theoretical question of what benefits explanations may offer to recommender systems.

- **Empirical data on the role of personalization in explanations.** We conducted a set of four experiments investigating what constitutes good content for an explanation. In particular, we focused on the role of personalization of explanations on effectiveness, persuasion, and satisfaction. We investigated a form of personalization that could be realistically achieved through scalable natural language generation (NLG), and using data available from a commercial e-commerce web service. We report the results for two distinct domains: movies and cameras. Since in these experiments we approximate trying the items (where users only read online reviews), we also conducted an experiment where users actually tried items.

- **Better understanding of the metric of effectiveness.** We used a metric for effectiveness based on a change of opinion: before and after trying an item. While this metric has previously been used by Bilgic and Mooney (2005), we adapted it during the course of a number of surveys and experiments. In this manner, we contribute to a deeper understanding of the metric, and its relation to the underlying data. We discuss when it is suitable to consider the signed value and when it is not. For example, when considering the average change of opinion for a number of items, it is better to use the absolute value to measure the aggregated error. Likewise, the existing metric did not give an indication of whether over or underestimation is preferable to users, or if this preference might be a domain dependent factor. In addition, we highlight the relation of the metric to the underlying data distribution.

1.3 Overview of the remainder of the thesis

This thesis can be divided into two main components: a theoretical component, and an empirical component. The first component (Chapters 2-4) aims to give a literature review of the field, and an idea of what explanations in recommender can be good for, and in particular how they may contribute to the effectiveness of a recommender system. The second component (Chapters 5-7) describes a series of experiments and studies aimed to help evaluate the added value of personalized explanations.

Chapter 2 - Recommender systems

In recent years there have been an increasing number of systems that help users find items that are relevant or interesting to them, so called recommender systems. Based on user input (implicit or explicit), preferences for unseen items (such as movies) is inferred. These systems can be seen as the next generation of expert systems, giving recommendations of items to try or buy rather than give expert advice. The underlying algorithm for selecting items in a recommender system may pose limitations or facilitate certain styles of explanations. For this reason, it is important to have at least a basic understanding of the algorithms, in order to be able to understand their effect on the explanations that can be generated (although we later in the thesis also note that there are cases where explanations do *not* reflect the underlying algorithm). Thus, in this chapter, we describe the most commonly used algorithms in recommender systems, and the strengths and weaknesses of each. We also discuss a number of ways in which recommender systems can be evaluated. This chapter is mostly meant as a review, and can be used as a reference.

Chapter 3 - Explanations in recommender systems

A recommended item can be described or justified to a user. For example, Amazon does this by saying something along the lines of: *“This item is recommended to you because...”*. This chapter offers an overview of explanations in several academic and commercial recommender systems, and discusses these in relation to expert systems. Based on a literature review, we begin by defining seven reasons for using explanations in terms of aims. This classification of explanatory aims is a novel contribution. We also make a clear distinction between different aims, such as between transparency (how was this recommendation made) and effectiveness (why would the user like or dislike this item). We claim that there is a difference between explanations of recommended items, and explanations of the recommendations themselves, and focus on the former. This chapter describes different ways of evaluating these explanatory aims, as well as previous work on evaluations of explanations in expert systems, and relates these evaluation metrics to the criteria by which recommender systems are normally measured (discussed in the preceding chapter). Since explanations are likely to be linked with the presentation of items, recommendation algorithm, and interaction with the system, these factors are also discussed in this chapter.

Chapter 4 - Effectiveness and personalization

In this chapter we go deeper into the definition and metrics for the aim of decision support. We raise the issue of potential trade-offs between persuasion and effectiveness. We review the previous literature, and consequently raise the question of whether using item features tailored to a user's preferences could aid decision support. In this chapter we also consider the severity of over- contra underestimation in various product domains, varied on two dimensions (high vs. low cost, and subjective vs. objective valuation).

Among other things, we saw that participants perceived (assumed explanations that caused) overestimation as less helpful than (assumed explanations that caused) underestimation, especially in high investment domains. We also found that *perceived effectiveness* (helpfulness) was affected by where on the scale a prediction error takes place.

Chapter 5 - Content for explanations: movies

Before we could study the role of explanations on effectiveness in any domain, we had to learn what helps users make decisions. Movies were selected as an initial domain, as data (e.g. ratings, properties, reviews) for movies are amply available. In this chapter, we describe a number of user studies including corpus analysis of online movie reviews, focus groups and a questionnaire. This resulted in a short list of relevant (domain specific) features, as well as an intuition that users differ in terms of what kind of information about a movie they consider important. Unfortunately this list was abbreviated once we realized which of these features were directly available via the commercial service we used later (see Appendix C on implementation), but was still used as a reference to guide the selection from the available options. This chapter also outlines a general methodology for eliciting relevant features that can be reused in another domain.

Chapter 6 - Personalization experiments

This chapter describes three experiments evaluating the effectiveness, persuasion, and satisfaction with explanations of recommendations in two domains: cameras and movies. All three experiments were based on a testbed system (described in Appendix C) using some of the features elicited in Chapter 5.

In all three experiments we compared explanations generated by our testbed with three degrees of personalization. Some of the results of the first experiment in the movie domain could have been due to confounding factors, and so we conducted a second in order to address this possibility. *For both experiments in the movie domain, we found that non-personalized explanations were more effective than personalized and baseline explanations. However, our participants were more satisfied with personalized explanations.* Knowing from Chapter 4 that high investment domains are more likely to be sensitive to overestimation, we repeated the experiments in a domain that differed to movies in respect to this dimension. *In this third experiment we found that non-personalized explanations were more effective than personalized and baseline explanations, but that our participants preferred personalized explanations.*

The design of the two experiments was nearly identical, and participants read online reviews rather than trying the items. As the item ratings on the website where they read the reviews are biased towards positive ratings, one explanation of our replicated results could be our design rather than the explanations. We investigated this in the next chapter.

Chapter 7 - Final personalization evaluation

In this chapter we evaluated true effectiveness rather than approximating it by letting participants read online reviews. In this experiment, we asked participants to actually try the items. Movies were considered easier to evaluate than cameras, and short movies were selected in order to decrease experiment duration. In addition, we had to carefully select the movies, partly due to ethical considerations, and partly so that they would contain relevant features such as known actors and directors.

This time, we found that the baseline explanations were most effective, and that participants were most satisfied with the non-personalized explanations. Baseline explanations also led to the largest amount of opt-outs. We believe that the reason the baseline was more effective this time is due more to the choice of materials and the nature of the baseline than the change in experimental design. An elaborate discussion of the result, and why we believe the material choice was decisive, is presented this chapter.

Chapter 8 - Conclusion and future Work

In this chapter we summarize our findings and ideas for future work. In particular, we discuss what our experiments imply with regard to the relevance and potential utility of explanations, as well as the role of personalization. We also reflect over the lessons we have learned about the metric we used for effectiveness, and highlight the effects underlying data could have on recommendation (as well as explanation) quality. In addition, we summarize our thoughts about presentational choices.

1.4 Related Publications

- **N Tintarev.** Explaining Recommendations. *Doctoral Consortium User Modeling'07*, pp. 470-474.
- **N Tintarev.** Explanations of Recommendations. *Doctoral Consortium Recommender Systems'07*, pp. 203-206
- **N Tintarev & J Masthoff.** Effective Explanations of Recommendations: User-Centered Design. *Recommender Systems'07*, pp. 153-156.
- **N Tintarev & J Masthoff.** A Survey of Explanations in Recommender Systems. In G Uchyigit (ed), *Workshop on Recommender Systems and Intelligent User Interfaces associated with ICDE'07,* pp. 801-810.
- **N Tintarev & J Masthoff.** Over- and underestimation in different product domains. *Workshop on Recommender Systems associated with ECAI, 08*, pp. 14-18.
- **N Tintarev & J Masthoff.** Personalizing Movie Explanations Using Commercial Meta-Data. *Adaptive Hypermedia'08*, pp. 204-213.
- **N Tintarev & J Masthoff.** Problems Experienced When Evaluating Effectiveness of Explanations for Recommender Systems, UMAP'09, pp. 54-63.
- **N Tintarev & J Masthoff.** Handbook Chapter on Explanations in Recommender Systems (under review 2009).

Chapter 2

Recommender systems

Recommender systems suggest items to purchase or examine based on users' preferences and interests. An early description of recommender systems was *" ...[a system where]... people provide recommendations as inputs, which the system then aggregates and directs to appropriate recipients"* (Resnick and Varian, 1997). The definition of recommender systems has grown broader since. Current recommender systems are more automated and create a user profile in order to propose a small selection of items out of a large variety of options. This profile can be based on a combination of *implicit* data, i.e. according to the user's patterns of use (e.g. Ardissono et al., 2004; Zimmerman et al., 2004) or, *explicit* data, where the user briefly, and throughout usage, specifies their preferences to the system (e.g. Billsus and Pazzani, 1999; McCarthy et al., 2004). For example, a system which sells books may recommend new books for a user to buy based on which books they have looked at or bought in the past (implicit rating), or how they have actively rated books (explicit rating).

Recommender systems can also differ by the extent to which they engage in a dialog with a user. In "*single-shot*" recommender systems, each user request is treated independently of previous ones. "*Conversational*" systems on the other hand are more interactive (Burke, 2002; Rafter and Smyth, 2005), and users elaborate their requirements over the course of an extended dialog. In particular, the user can supply feedback on the recommended items which influences the next set of recommendations. A discussion of different feedback styles can be found in McGinty and Smyth (2002); Tintarev and Masthoff (2007), as well as in Section 3.7.

These personalized recommender systems have become valuable tools for sifting through large amounts of data. Criteria such as retrieving "individualized" as well as "interesting and useful" content have become particularly important (Burke, 2002). In other words, it is as important that recommendations are interesting, as it is for them to be accurate. If this is done well, the system may gain increased usage and loyalty (e.g. Rashid et al., 2002; McNee et al., 2003b).

Application domains for recommender systems are widely distributed. Previous systems exist in domains such as movies (e.g. Ziegler et al., 2005), music (e.g. *pandora.com, last.fm*), books (whichbook.net, librarything.com), electronic program guides (e.g. O'Sullivan et al., 2004; Ardissono et al., 2004), digital cameras (e.g. Reilly et al., 2004c), computers (e.g. Reilly et al., 2004a) and holidays (e.g. McSherry, 2005)[1].

In Section 2.1 we discuss commonly used recommendation algorithms, in Section 2.2 the trade-offs involved in using different algorithms, and in Section 2.3 the criteria by which they have been evaluated. Finally, we conclude with a short summary in Section 2.4.

2.1 Recommendation Techniques

Modern recommender systems use a number of methods; Table 2.1 summarizes common techniques and their respective background data, input, and process (ouput). This chapter discusses the following techniques: *demographic-based filters* (e.g. Ardissono et al., 2004), *content-based filters* (e.g. Ferman et al., 2002), *collaborative filters* (e.g. Rashid et al., 2002), *knowledge-based* (e.g. McCarthy et al., 2004; Burke, 2002), and *utility-based filters* (e.g. McSherry, 2005; Reilly et al., 2004b). Content-based and collaborative-based filters are the most common types of recommendation algorithms. This is because they are based on rating data, which is relatively easy to collect, and for which there are already established data-sets. Consequently, this chapter will discuss content-based and collaborative-based algorithms in most detail.

Many recommender systems combine different algorithms in hybrid systems (e.g. Ardissono et al., 2004; Pazzani, 2002; Ziegler et al., 2005) to counterbalance the weaknesses of the individual methods. In the next section we describe these methods, give examples, and discuss their respective strengths and weakness. We conclude this section on algorithms with a brief discussion of the state of the art.

2.1.1 Collaborative filters

Description

A recommender system may use correlations between users as a basis for forming predicted ratings of recommended items. That is, recommendations for unseen items are based on information known about ratings of other users, and how these ratings correlate with the ratings of the current user. This approach is called *collaborative filtering* (in places abbreviated as *CF*) (Ziegler et al., 2005). Collaborative filtering can be either

[1] Websites retrieved June 2007.

Table 2.1: Recommendation techniques, modified from Burke (2002). **U** is the set of users whose preferences are known, $\mathbf{u} \in U$ is the user for whom recommendations need to be generated, and $\mathbf{i} \in \mathbf{I}$ is an item for which we would like to predict u's preferences.

Technique	**Background**	**Input**	**Process**
Collaborative	Ratings from **U** (for a sub-set) of items in **I**.	**u**'s ratings of items in **I**.	Identify users that are similar in ratings to **u**, and extrapolate from their ratings of **i**.
Content-based	Features of items in **I**	**u**'s ratings (for a sub-set) of items in **I**	Generate a classifier that fits **u**'s rating behavior and use it on **i**.
Knowledge-based	Features of items in **I**. Knowledge of how these items meet **u**'s needs.	A description of **u**'s needs or interests.	Infer a match between **i** and **u**'s need.
Utility-based	Features of items in **I**.	A utility function over items in **I** that describes **u**'s preferences.	Apply the function to the items and determine **i**'s rank.
Demographic	Demographic information about **U** and their ratings of items in **I**.	Demographic information about **u**.	Identify users that are demographically similar to **u**, and extrapolate from their ratings of **i**.

heuristic-based (also called memory-based) or model-based (Adomavicius and Alexander Tuzhilin, 2005; Breese et al., 1998). Heuristic-based algorithms make predictions based on all the ratings given by the users. Model-based algorithms in contrast, use a subset of ratings to build a model, such as a naive Bayesian classifier, which is subsequently used to make rating predictions.

We will use a heuristic-based algorithm as an illustration, and refer to (Adomavicius and Alexander Tuzhilin, 2005; Breese et al., 1998) for an overview of alternative collaborative algorithms, and model-based algorithms in particular.

How it works

The most common CF approach is *user-based*, which uses correlations between a user **u** and other users, in order to decide the extent to which these users should impact on the recommendations for this user. This method largely presupposes that recommended items can be given a numerical value or rating of some sort by the users. The recommendation

Table 2.2: Example for collaborative filtering

	Adam	Brian	Carl	Diana	Eve
Starwars	-	+	+	+	-
Pretty Woman	+	+	+	-	+
101 Dalmatians	+	-	+	-	+
Terminator	-	+	-	+	-
Little Mermaid	+	-	+	+	?

algorithm can be broken down into the following steps:

1. Establish how similar other users are to the user **u**.
2. Use this similarity to weight the ratings of these users for the recommended item **i** in a weighted sum.
3. Apply additional filters or weights.

To gain an intuition of how this works, we survey Table 2.2. We see that Adam and Eve have previously rated movies in the same way, so this is likely to influence Eve's prediction for the Little Mermaid positively. We can also see that Eve and Diana tend to disagree in their ratings, again increasing the likelihood that Eve will rate the Little Mermaid highly.

An *item-based CF* recommender system (Ziegler et al., 2005; Rashid et al., 2002), rather than basing recommendations on the similarity between users, bases them on similarity between items. The recommendation algorithm can be seen to be broken down into the following steps:

1. Establish how similar items are to item **i** (with regard to rating patterns).
2. Use this similarity to weight the average rating (over all users) for these items to predict the rating for the recommended item **i** in a weighted sum.
3. Apply additional filters or weights.

Again, to gain an intuitive idea of how the algorithm works we revisit Table 2.2. We can see that users who rated 101 Dalmatians give a similar rating to the Little Mermaid. The average rating of 101 Dalmatians is therefore given a high weight relative to other movies when forming a prediction for the Little Mermaid for new users.

The techniques for user and item-based algorithms are similar, so in the next paragraph we describe the three steps in terms of user-based CF only. *Similarity* between users is often computed by creating vectors of the user's ratings, and comparing them to other users using Pearson's correlation (Resnick et al., 1994) or cosine distance (Salton and McGil, 1986) etc.

After computing a degree of similarity between the current (active) user and other users, the system *predicts a rating* for a given item (or items). Alternatively, only most similar users (neighbors) are used for computing the predicted rating(s). The technique is therefore sometimes also called K nearest neighbors (or K-nn)

2.1.2 Content-based filters

Description

Content-based recommender systems base recommendations on user ratings and similarity between *items*. That is, while CF filters are based on correlations between users, content-based filters are based on correlations between items. Although this is not the only way to represent items, they are commonly represented as sets of terms or keywords that can be given relative importance using weights. Similarly to collaborative filtering, content-based filtering may use Pearson's correlation (Resnick et al., 1994) or cosine distance (Salton and McGil, 1986) to measure the similarity between two items based on their keywords. For cosine similarity, the vectors describe the (weighted) frequency of the keywords, or terms, in it. Each term defines a direction, or rather a dimension, in the vector-space. Similarity is then a measure of proximity between these vectors. For Pearson correlation, the similarity of two items is computed as a weighted sum of all the keywords.

Naturally, other types of similarity between items are possible. For example, similarity may be semantically defined, or based on other techniques such as image similarity (Jacobs et al., 1995) [2].

Table 2.3: Example for content-based filtering

	Action	Sci-fi	Comedy	Romance	Children	Eve
Starwars	Y	Y				-
Pretty Woman			Y	Y		+
101 Dalmatians					Y	+
Terminator	Y					-
Little Mermaid					Y	?

[2] See also Retrivr: http://labs.systemone.at/retrievr/, retrieved June 2008

How it works

To gain an intuition of how content-based filtering may work, we survey Table 2.3. Each movie has several keywords (in this case genres). We see for example that Eve does not like movies with the keyword Action, and that she did like the other movie with the keyword Children. The weight for Action would be negative, and the weight for Children would be positive. In this case, the movie only has the term Children, and so it is given a high score and is recommended. If the movie had also belonged to the genre Action, the predicted rating of the movie would be lower if it just belonged to the genre Children. How much lower would depend on the relative weighting of this keyword.

Content-based filtering also allows for additional weighting. In many systems the weights are based on how informative the term is for an item compared to other terms, as well as how much this term discriminates this item from other items. Commonly occurring keywords are given lower weights; capturing the intuition that common words are not as useful in identifying the topic of an item, while words that occur less frequently are more indicative of a topic. This method is commonly called TF-IDF (term-frequency, inverse-document-frequency) (Salton and McGil, 1986).

The filtering can also occur before the similarity metric is applied. For example, the system might only use the top x most frequent keywords when calculating similarity between two items. TF-IDF requires x to be preset. An alternative is the Winnow algorithm (Lewis et al., 1996) which dynamically identifies a suitable number of keywords, converging on a set of weights that typically assign high weights to a small percentage of the words.

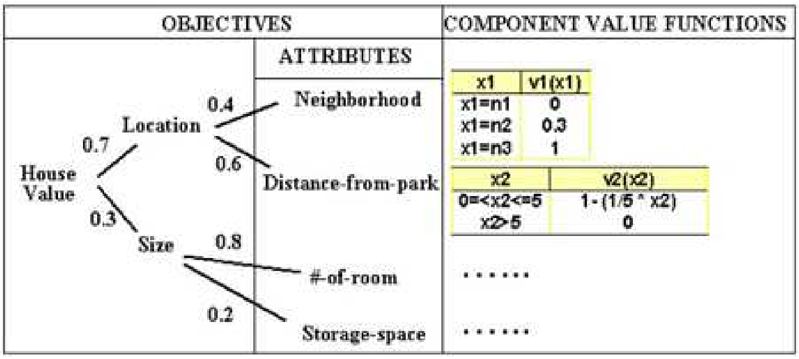
2.1.3 Knowledge-based and Utility-based filters

Description

Both knowledge-based and utility-based filters can be seen as particular types of content-based filters. In other words, item properties (such as price) are used in order to make recommendations. They are however worthy of their own section as these types of systems make a more explicit connection between user requirements and available options than content-based filtering.

We choose to combine the sections for knowledge- and utility-based filters because utility-based filters can in addition be seen as a type of knowledge-based filter. Both filters attempt to make a matching between a user's needs and the available options. The difference between the two is that while utility-based systems require users to do their own mapping between their needs and features of products, knowledge-based systems have deeper domain knowledge. In particular knowledge-based systems are able to reason about how a particular item meets a particular user need. For example, it can combine item properties into dimensions that are more comprehensible to a user, or how they might interact. For example, a cheaper camera will probably have lower resolution.

How it works



x1	v1(x1)
x1=n1	0
x1=n2	0.3
x1=n3	1

x2	v2(x2)
0=<x2<=5	1-(1/5 * x2)
x2>5	0

Figure 2.1: Sample additive multi-attribute value function, Carenini and Moore (2000b)

These types of filters recommend items to users based on which properties the users find important. For example, Carenini and Moore (2000b) use methods from decision support such as Multi-Attribute Utility Theory (MAUT) (Clemen, 1996) to weigh item properties in relation to a user as well as the relative importance between properties. They use an additive multi-attribute value function (AMVF) to model user's preferences with respect to items (in their case houses).

Each user's preferences is represented as value tree, and a set of component value functions, one for each attribute of the item. This is illustrated in Figure 2.1 for a house. The value tree represents the properties of a house, and the arcs of the tree are weighted (sum of all arcs is 1) to represent the importance of the value of an item feature in contributing to the value of its ancestors (parents, grandparents etc) in the tree. In the Figure 2.1 location is more than twice as important as size in determining the value of a house. The component value functions represent how important item features are for a user (value from 0 to 1). For instance, neighborhood n2 has a preference value of 0.3, and a distance-from-park of 1 mile has a preference value of $(1 - (1/5 * 1)) = 0.8$.

The value of a house is thus a weighted sum of all the arcs in the value tree for the item, and the component value functions for a given user. In this way, it is possible to compute how suitable the item is to the user, and how valuable any feature of that house is for that person. So, in the example above houses near parks are likely to be considered more valuable than those further away, and the size is likely to be less decisive than location.

While some knowledge-based systems are one-shot, more recent systems learn utility for item properties over time, and even adapt to changes in user preferences. In particular we discuss systems which use a method called critiquing.

Critiquing

Critiquing uses a constraint based form of feedback, which takes into account that it may be easier for a user to ask for alterations to recommended items rather than construct a query. The user is presented with an item and gives feedback on a feature of that particular item such as its price, e.g. “like this but cheaper”. This method has been extended in a number of ways. *Incremental critiquing* allows the user to successively apply constraints (Reilly et al., 2004c), or critiques. Another extension called *Dynamic critiquing* presents users with compound critiques (Reilly et al., 2004b). *Compound critiques* are defined by the authors as a combination of critiques such as “Less Memory, Lower Resolution and Cheaper” in the camera domain. These compound critiques are selected to reflect the remaining product options (Reilly et al., 2004b). Compound critiques are computed using the Apriori algorithm (Agrawal et al., 1996), which finds the types of combinations of features that occur in a dataset. McCarthy et al. (2005) suggest that using rules with low support (uncommon that $propertiesA \rightarrow propertiesB$) are likely to help generate diverse critiques and help users find the item that they want quicker as they function as discriminators.
(McSherry and Aha, 2007) suggest a more knowledge intense approach called *Progressive Critiquing* primarily aimed at alerting the user to the possibility that none of the available products may be acceptable. (McSherry and Aha, 2007) also suggest how implicit relaxation of critiques can be used when users over-critique i.e. are overly stringent in their requests.

Critiquing is particularly useful in terms of how well it helps the system explain recommendations, as demonstrated in e.g. McCarthy et al. (2004); McSherry (2005).

2.1.4 Demographic-based filters

Description

Demographic-based filters (also called stereotype-based filters) use known user characteristics in order to classify users and model user preferences into classes. A recommender system can define a typical user according to their socio-demographic information, such as postcode, age or gender. Similarly, a combination of demographics can be used to define a stereotype. This stereotype assumes an interest profile, such as particular genres in the movie domain. The user’s interests are subsequently classified with regard to how strongly they fit the stereotype, or rather how well they fit *several* stereotypes.

One could say that demographic-based filters work similarly to collaborative-filters in that they derive similarities between users, but use different types of data. Here similarity is based on demographic information of users rather than their rating patterns.

How it works

As an example, we describe a system which has been used in the domain of personalized program guides (television) described in Ardissono et al. (2004). This system estimates users' preferences in two steps. In the first step, the user is matched against a number of stereotypes using the available socio-demographic and interest data. The result is a degree of matching with respect to each stereotype. In the second step, the users' preferences are estimated by combining the *predictions* of each stereotype, weighted according to the degree of matching with the user.

Other examples are the InfoFinder Agent (Krulwich, 1997) which retrieves relevant documents based on demographic groups from marketing research, and Billsus and Pazzani (1999) who use machine learning to build a classifier of news interests based on demographic data.

2.1.5 The "state of the art" and the Netflix prize

In 2006, the American online movie company Netflix released a dataset containing 100 million movie ratings and challenged the machine learning and computer science communities to improve the accuracy of its recommendation engine, Cinematch [3]. At the time of writing (May 2008) this competition is approaching the end of its second year, and the learned lessons have been shared within the research community.

There seems to be a consensus that the Netflix problem is best tackled by a combination of matrix factorization (a model-based approach) and the K-nearest neighbors (see Section 2.1.1 for a description) approach (Takács et al., 2007). The team who won the first Netflix progress prize after the first year (i.e. Bell and Koren, 2007), used a combination of matrix factorization and a modified K-nearest neighbors approach.

Bell and Koren (2007) argue that while k-NN approaches are effective at detecting localized relations (e.g. within clusters of movies), they are poor at capturing weaker signals encompassed in *all* of a user's ratings. Latent factor models (based on matrix factorization) on the other hand are effective at estimating overall structure that relates simultaneously to most or all movies, but they are poor at detecting strong associations among a small set of closely related movies.

Bell and Koren (2007) also highlight the importance of making use of *which* movies users rate regardless of *how* they rated these movies for matrix factorization. A similar

[3] http://www.netflixprize.com

result was previously cited by Bridge and Kelly (2006) for collaborative filtering.

2.1.6 Hybrids

Many recommender systems use a combination, or hybrid, of methods to counter-balance the weaknesses in each algorithm. In the next sections we will discuss the trade-offs involved with each algorithm in Section 2.2, and return to how hybrid solutions can be used to make best use of respective strengths in Section 2.2.9.

The most common type of hybrid is collaborative/content (e.g. Balabanovic and Shoham, 1997; Basu et al., 1998; Melville et al., 2002; O'Sullivan et al., 2004; Symeonidis et al., 2007). As previously mentioned, this is largely due to the availability of rating data. Other solutions have combined demographic user classes and content-based filters using implicit behavior and explicit preferences (Ardissono et al., 2004), collaborative filtering and demographic (Pazzani, 1999) or collaborative filtering and knowledge-based filters (Towle and Quinn, 2000).

Burke (2002) discusses different types of hybrids used, and discusses how algorithms can be combined. Some of the methods are order independent, while others give different recommendations based on which algorithm is applied first. For example, weighted methods are not sensitive to order. In weighted hybrids the scores (or votes) of several recommendation techniques are given relative importance (a weight) and combined together to produce a single recommendation. Other hybrids, such as cascades are sensitive to order. In these hybrids, one recommender *refines* the recommendations given by another.

2.2 Trade-offs of different methods

In this section we discuss the strengths and weakness of different methods. Table 2.4 compares the five different recommendation algorithms (collaborative, content, utility, knowledge and demographic-based filtering). The trade-offs are assigned capital letters to which we refer in the following subsections.

2.2.1 "Cold-start"

Recommender systems suffer from two types of *cold-start* problems, i.e. situations where they do not have enough information, **(I)** and **(J)** in Table 2.4. The first has to do with *new users* (I), and the second with *new items* (J).

Often a recommender system does not have enough information about a new user in order to deduce anything about their taste. The new user cold-start problem occurs in systems with collaborative-based, content-based, and demographic filtering. In these types of system, a new user needs to supply information about themselves or their preferences

Table 2.4: Trade-offs between recommendation techniques, adapted from Burke (2002)

Technique	Strengths	Weaknesses
Collaborative filtering (CF)	A. Can identify cross-genre niches B. Deep domain knowledge not needed C. Adaptive: quality improves over time D. Implicit feedback sufficient	I. New user cold-start J. New item cold-start K. ”Gray sheep ” problem L. Quality dependent on large historical dataset (sparsity) M. Stability vs. plasticity problem
Content-based (CN)	B,C,D	I, L, M
Utility-based (UT)	E. No cold start. F. Sensitive to changes of preference G. Can include non-product features	N. User must input utility function O. Suggestion ability static (does not learn)
Knowledge-based (KB)	E,F,G H. Can map from user needs to products	O. P. Knowledge engineering required.
Demographic (DM)	A,B,C	I,K,L,M Q. Must gather demographic information

in order to create a basis for future recommendation. Knowledge-based filters use deep domain knowledge circumventing this problem, while utility-based filters have a related problem requiring the user to input a utility function (**N**).

Another type of cold-start regards new items. This problem occurs in collaborative filtering systems only. Initial ratings for each item (preferably by a large number of users) is necessary in order to make recommendations about it, but when a new item enters a system it has no or little ratings and is not included in neighborhood formation or similarity comparisons. On the other hand, other types of systems may require manual annotation of item features as they enter the system.

2.2.2 Type of input

While content-based and collaborative-based filters can manage with implicit ratings from users **(D)**, such as viewing time for video media, knowledge and utility-based, and demographic filters are often limited to explicit input from users. Knowledge and utility-based systems require input of user preferences and the utility function **(N)** respectively, while demographic filters need to learn more about the user's demographics **(Q)**. In the age of online social networks it is arguable that demographic information can be mined, but even in this case explicit consent should be given by the user.

2.2.3 The *"portfolio effect"* and cross-genre niches

Content-based recommender systems often suffer from uniform recommendations. For example, Amazon's recommender system initially suffered from a "portfolio effect" (Linden et al., 2003; Burke, 2002), i.e. offered recommendations so similar they were of little to no use to the user. Nor did they inspire to explore new avenues, i.e recommendations lacked serendipity. Such a recommender might for example continuously recommend books by the same author, or even different versions of the same book even when the user has already bought it. This highlights the care that needs to be taken when selecting similarity metrics, and the need for diversity within a set a recommendations.

In contrast, collaborative filtering allows for more diverse recommendations than the other recommender algorithms (except possibly demographic filters which can be seen as a subtype of collaborative filters). CF systems help users discover items that are similar, but not identical, and discover items or groups of items they may not have otherwise considered in cross-genre niches **(A)**.

2.2.4 "Sparsity"

CF systems in particular are dependent on having a sufficient number of item ratings per user, as they require a high level of overlap between users in order for them to be considered neighbors, especially if the total number of items and users is large. This is often called the *sparsity problem*. Item-based CF has been shown to significantly enhance the performance of a recommender system in sparse profile spaces compared to user-based CF (O'Sullivan et al., 2004).

Similarly, content-based and demographic filters require a large historical data-set in order to learn enough about a user to make strong recommendations **(L)**.

2.2.5 “Gray sheep”

While they do not suffer from the portfolio effect as do content-based filters, CF and demographic filters suffer from the “gray sheep” phenomenon **(K)**, where the recommender system has trouble recommending items to users with unusual preferences. In CF systems, the ratings of an ‘unusual’ user will not correlate well with the ratings of other users. This can be seen as a particular form of sparsity, as it can be hard to sample the entire search space and items may form clusters which do not mix between groups of users (Rashid et al., 2002).

Demographic filters, like collaborative-filtering, can identify cross-genre niches, helping users to discover new types of items **(A)**. However, they also suffer from their own type of “gray sheep” phenomenon. It can be hard for a demographic filtering recommender system to discover and define all possible stereotypes. They can capture similarities between users, but users may not fit neatly into a stereotype, or even a combination of stereotypes. Demographic filters also require gathering of demographic information **(Q)**.

2.2.6 “Starvation” and “Shilling”

In the same way as an unusual user can be “starved” to the benefit of the interests of “common” users (i.e. the “gray sheep phenomenon”), items can be starved too (Krause, 2006). In CF systems, popular items become easier to find as more users rate them, at the expense of unpopular or undiscovered items which become more difficult to discover (Rashid et al., 2002; McNee et al., 2003b). A large dataset alleviates this problem **(L)**, but may not solve it altogether. This also makes a CF recommender system susceptible to malicious attacks, so called *shilling*, such as injections of ratings which can cause the popularity of an item to either sink or soar (Mehta, 2007; Mobasher et al., 2006).

In a similar manner, some items in the other types of systems may be difficult to retrieve (e.g. requiring a complex utility function in a utility-based system **(N)**), making them inaccessible to most users. Making sure that all, or most items are available for retrieval is an additional requirement involved in knowledge engineering for knowledge-based systems **(P)**.

2.2.7 Adaptivity and sensitivity to change

As previously mentioned, the quality of recommendations from content, collaborative and demographic based filters improves over time **(C)**, making it dependent on a large historical dataset **(L)**.

Recommender systems which establish user preferences over time may also become rigid. Once a preference has been established for a user, it becomes very hard to change.

This is not the case for knowledge-based recommender systems where a user's requests immediately modify preferences, and where the preferences are not stored for long periods of time **(F,M)**.

On the other hand, for knowledge-based systems, once a utility function is found, the possible range of recommendations stays static and does not improve over time **(O)**. This is not the case for content, collaborative and demographic filtering. Likewise, the range of recommendations is limited to that of the user designed utility function (though this may increase throughout usage the system (Reilly et al., 2004c)) or the knowledge of system designers.

2.2.8 Knowledge engineering and utility functions

Knowledge-based systems do not suffer from the above problems. They do not require a long learning process **(E)**, and are sensitive to change of user preferences **(F)**. On the other hand they have their own distinct limitations.

Knowledge-based systems may be able to map user needs to products **(H)**, and include non-product features in the reasoning **(G)**. This requires knowledge-based systems to have domain specific knowledge **(P)**, unlike the other common recommender algorithms **(B)**. This is however a broad generalization, as for example content-based systems rely on descriptive data which is often domain dependent. In contrast knowledge-based systems require deep knowledge engineering, and utility-based systems require users to input their own, and often complex, utility function **(N)**.

2.2.9 Trade-offs and hybrids

The most common type of hybrid of collaborative and content-based filtering helps circumvent the sparsity and item cold start-problems (Balabanovic and Shoham, 1997; Basu et al., 1998; Melville et al., 2002; O'Sullivan et al., 2004; Symeonidis et al., 2007). The advantage of content-based filtering over collaborative filtering, is that it does not suffer from the cold start problem for new items. Nor does it suffer from the sparsity problem. The recommendation of a new item is based on its similarity with other items one user has rated, not overlap with other users. Collaborative filtering on the other hand is not dependent on descriptive data about items. It offers recommendations based on similarity between users and user ratings only. This means that recommended items might not have an easily identifiable similarity, but suffer less from uniformity, offering more serendipitous recommendations. While these hybrids are better than the two individual methods used on their own, they still suffer from many disadvantages such as new user cold-start, "gray sheep" and lack of plasticity.

Hybrids of demographic and content-based filters allow the recommender system a basis on which to form a user profile, thus resolving cold-start for new users (Ardissono et al., 2004). However, demographic filters require a great amount of demographic knowledge.

Hybrids combining collaborative filtering with demographic (Pazzani, 1999) or knowledge-based (Towle and Quinn, 2000) filters both alleviate the new item problem while maintaining the ability to improve recommendations with time and give cross-genre recommendations. Knowledge-based filters can also make recommendations with few and sparse users where CF cannot, but they require more deep domain knowledge.

2.3 Evaluating recommender systems

It is important that recommended items are both *accurate* (measured by e.g. precision and novelty) and *useful* (measured by e.g. novelty). It has been found that users of recommender systems are not always confident that the system will identify their preferences correctly (Pittarello, 2004). It is important that a recommender system can identify which items may be important to a user, and which ones probably are not. Accuracy is not everything however, as accurate recommendations are not necessarily useful. That is, an accurate recommendation may not be consumed by a user and/or help them (McNee et al., 2006b). The usefulness of recommendations can also be influenced other factors. For example, the spread or *diversity* of a recommendation list can affect users' opinion of a system. Users may be concerned that simply following recommendations would lead them to miss some important novel item (Pittarello, 2004). *Coverage* regards the range of items that potentially could be recommended to a user, and is therefore an important factor for both recommendation accuracy and usefulness. We discuss these and other evaluation factors in the sections below.

2.3.1 Accuracy

Firstly, it is important for recommendations to be accurate. *Accuracy* metrics are often used in information retrieval as the proportion of correctly classified items (Billsus and Pazzani, 1999). Depending on the application, one accuracy measure may be more relevant than another. For example, precision could be considered more important than recall in the news domain. Perhaps it is more important to supply relevant news items (precision), than to supply all 100 items regarding a single topic (recall). Notwithstanding, both precision and recall are relevant measures of accuracy (Maybury et al., 2004) and should be measured together so that the system does not do very well on one measure but not the other.

Precision . This metric is an inverse measure of false hit rate, i.e. the ability to retrieve only relevant items (Billsus and Pazzani, 1999). For 'y' retrieved relevant number of movies, and 'n' the number of retrieved non-relevant movies, precision can be defined as:

$$P = \frac{y}{y+n} \tag{2.1}$$

Let us take an example in the movie domain. Let us assume a search, initiated by either user or the system, for movies staring Sean Connery. Low precision would imply that many of the clips found did not star Sean Connery, e.g. Bond movies with other actors. High precision, on the other hand, would mean that many of the retrieved movies would star Sean Connery.

Recall . This metric is also called hit rate, i.e. the ability to retrieve all the relevant items (Merlino et al., 1997). Using our previous example, high recall would be the ability to show *all* of the movies staring Sean Connery from the available dataset.

For 'y' retrieved relevant items, and 'm' missed relevant items, recall can be defined as:

$$R = \frac{y}{y+m} \tag{2.2}$$

F-score . It is easy to optimize for recall, but have low precision recommendations and vice versa. For example, a recommender system can return all the items in its catalog. All the relevant items will be included in this list (recall), but there will be irrelevant recommendations too (precision). This metric considers this trade-off and returns a weighted combination of the two previous metrics:

$$F = \frac{2 * (P * R)}{P + R} \tag{2.3}$$

MAE . If recommended items are evaluated in terms of numeric values, accuracy can be a measure of how close predicted ratings come to true user ratings. Mean average error (MAE) for a user *u* can be defined with predictions $p_u(item_k)$, and ratings $r_u(item_k)$, for sets $Items_u$ of products ($item_k \in Items_u$). This is a weighted difference between the prediction and rating, divided by the number of items rated $|Items_u|$:

$$MAE = \sum_{item_k \in Items_u} \frac{|r_u(item_k) - p_u(item_k)|}{|Items_u|} \tag{2.4}$$

ROC . The Receiver Operating Characteristics Analysis curve is a signal processing measure. It plots *recall* against $1-$ specificity, with specificity defined as the probability of a non-relevant item being rejected for recommendation (Swets, 1988). Points on the plotted curve then represent trade-offs supported by the filter, between recall and a metric which is similar to precision.

2.3.2 Coverage

Another neighboring concept is *coverage*, defined as the range of items in the domain for which predictions can be made. Commonly (e.g. Good et al., 1999; Herlocker et al., 1999; Ziegler et al., 2005) this is measured as the total of numbers of items for which predictions can be made as a percentage of the total number of items. A system that is too selective in recommendations will be disfavored by users who are afraid to miss novel items. In selection a system should, if at all possible, consider all possible options, and ensure that access to a wider range of items is not restricted.

It is important to highlight that coverage and accuracy should be measured together. The worse-case scenario illustrates this point: random predictions can give complete coverage but will have low accuracy.

2.3.3 Learning Rate

For recommender systems that are dependent on rich data and/or item ratings (e.g. content, collaborative and demographic based) it can be interesting to see how quickly they can give accurate recommendations. One can imagine three types of learning rates, overall learning rate, per item learning rate, and per user learning rate. The *overall* learning rate is recommendation quality as a function of the overall number of ratings in the system (or the overall number of users in the system). The *per-item* learning rate is the quality of predictions for an item as a function of the number of ratings available for that item. Similarly the *per-user* learning rate is the quality of the recommendations for a user as a function of the number of ratings that user has contributed. Most commonly learning rates are compared in graphs of quality (usually accuracy) versus the number of ratings (Herlocker et al., 2004).

It may also be important to measure how quickly a recommender system can adapt to a change in user preferences.

2.3.4 Novelty and Serendipity

Novelty . Users are likely to appreciate items that are familiar or similar to what they already know. This has been suggested for items in the news domain where *consistency* has been found to improve user satisfaction (Bell, 2002). Affirming recommendations can be used to establish rapport with new users (McNee et al., 2006b). In contrast *novelty*, the ability to retrieve new items, is also a positive contributor to users' satisfaction (Bell, 2002). Even though a certain item might match a user's preferences perfectly, the user will not be interested in it if it is too similar to previously recommended items (see also Section 2.2.3 on the "portfolio effect"). Novel items may be anticipated to a greater or lesser degree - a movie may be unseen, but very similar to other movies the user knows about. E.g. If the user has seen the first part of the "Lord of the Rings" trilogy, they probably know about the second part too. A balance needs to be struck between the known and the unknown. This is why in Billsus and Pazzani (1999) a user was able to specify whether they thought a news item is interesting or not, if they wanted to know more, or if they felt that they had already heard about the item previously . Novelty also reflects on accuracy. A recommender system which does not recognize different versions of an item as identical, and decides they are novel may present a user with them multiple times. If this item yielded perfect accuracy the first time, it might not do so for every subsequent exposure, as the user may not appreciate this redundancy.

Serendipity . This is a concept that is related to novelty, but is not altogether identical. Defined by Webster's Dictionary as *"the faculty or phenomenon of finding valuable or agreeable things not sought for"*. Some of the literature confounds novelty and serendipity. A "serendipitous" item is something completely unexpected, possibly an item a user would not search for on their own. Novel items on the other hand might be new, but not unexpected. Although serendipitous recommendations are not necessarily correct or accurate, they may introduce the user to interesting items they would otherwise not have discovered. Likewise, this may help the recommender system learn more about a user's interests. Current systems applying collaborative filtering (see Section 2.1.1 for more detail) are able to make such serendipitous recommendations.

2.3.5 Confidence

A recommendation can often be described by two values. Firstly, the strength of the recommendation, i.e. how much the recommender system predicts a user will like or dislike an item. Secondly, the confidence in this prediction, i.e. how strong the backing evidence is.

Herlocker et al. (2004) highlights the importance of differentiating the two: very extreme (either high or low) predictions are often made on the basis of small amounts of data, while predictions become less extreme as confidence increases over time. I.e. high strength may often mean low confidence, although this is not necessarily the case. The converse it also not necessarily true, low confidence does not necessarily mean high ratings and neutral ratings do not by default suggest high confidence.

Awareness of this differentiation can be used to the advantage of system designers.

For example, a recommender may choose to be risk-taking and generate recommendations for obscure, under-represented, or possibly unrelated items (McNee et al., 2006b). Even if it seems like a user would rate an item lowly, or if the system cannot accurately predict ratings for an item, a risk-taking recommender would still recommend this item.

2.4 Summary

In this chapter we have described a number of recommender system algorithms: content, collaborative, utility, knowledge and demographic-based algorithms. We have discussed the trade-offs involved with each recommendation algorithm, and how these trade-offs have been alleviated by hybrids. A brief section is also dedicated to the state of the art in the context of the Netflix prize.

We also discussed different ways of evaluating recommender systems such as accuracy, coverage, learning rate, novelty and serendipity, as well as confidence. In the next chapter (Chapter 3) we discuss explanations in the context of recommender systems. In Section 3.5 we define different roles explanations can play in recommender systems, and how these relate to evaluation metrics, and in Section 3.8 how the underlying algorithm can affect explanations.

Chapter 3

Explanations in recommender systems

3.1 Introduction

In recent years, there has been an increased interest in more user-centered evaluation metrics for recommender systems such as those mentioned in (McNee et al., 2006a). It has also been recognized that many recommender systems functioned as black boxes, providing no transparency into the working of the recommendation process, nor offering any additional information to accompany the recommendations beyond the recommendations themselves (Herlocker et al., 2000).

The definition that is used for explanations in this thesis primarily follows the first sense from the concise Oxford dictionary: *"to make clear by giving a detailed description"*. While the explanation may not be as detailed as for example a review, it should offer the user a sufficient enough description as to help them understand the item well enough to decide whether this item is relevant to them or not. We also consider the second sense suggested by the concise Oxford dictionary *"give a reason or justification for"*. In this case, the explanation aims to justify recommending the item. However, the recommendation itself can be assumed to be implicit through the mere presentation of the item. In addition, we include scenarios where a user may be presented with items that they do not like. Rather than omitting these items from presentation, the explanations give a justification for how the user should relate to the item, although they may well be negative in these cases.

Explanations can provide transparency, exposing the reasoning and data behind a recommendation. This is the case with some of the explanations hosted on Amazon, such as: *"Customers Who Bought This Item Also Bought ... "*. Explanations can also serve other aims such as helping users to make decisions about the items (effectiveness). In this way, we distinguish between different explanation criteria such as e.g. explaining the way the recommendation engine works (transparency), and explaining why the user may or may not want to try an item (effectiveness). An effective explanation may be formulated along the lines of *"You might (not) like Item A because..."*. In contrast to the Amazon example above, this explanation does not *necessarily* describe how the recommendation was selected - in which case it is not transparent.

This chapter offers guidelines for designing and evaluating explanations in recommender systems as summarized in Section 3.2. Up until now there has been little consensus as to how to evaluate explanations, or why to explain at all. In Section 3.3, we list seven explanatory criteria (including, but not limited to transparency and effectiveness), and describe how these have been measured in previous systems. These criteria can also be understood as advantages that explanations may offer to recommender systems, answering the question of *why* to explain. In the examples for effective and transparent explanations above, we saw that the two evaluation criteria could be mutually exclusive. In this section, we will describe cases where explanation criteria could be considered contradictory, and cases where they could be considered complementary.

We note already here that in the remainder of the thesis we will primarily focus on the criterion of effectiveness, as the work on this particular criterion has been limited. In any evaluation of a given criterion it is however important to realize that this is one of many possible criteria, and it is worthwhile to study the trade-offs involved with optimizing on a single criterion. We will take this approach of considering multiple criteria in our evaluations in Chapters 6 and 7.

Expert systems can be said to be the predecessors of recommender systems. Therefore, in Section 3.4 we therefore briefly relate research on evaluating explanations in expert systems to evaluations of explanations in recommender systems in terms of the identified criteria. We also identify the developments in recommender systems which may have caused a revived interest in explanation research since the days of expert systems.

Additionally, explanations are not decoupled from recommendations themselves or the way in which users can interact with recommendations. In Section 3.5, we consider that the underlying recommender system affects the evaluation of explanations, and discuss this in terms of the evaluation metrics normally used for recommender systems (e.g. accuracy and coverage). We mention and illustrate examples of explanations throughout the chapter, and offer an aggregated list of examples in commercial and academic recommender systems in Table 3.4. We will see that explanations have been presented in various forms, using both text and graphics. In Section 3.6 we mention different ways of presenting recommendations and their effect on explanations. Section 3.7 describes how users can interact and give input to a recommender system, and how this affects explanations. Both are factors that need to be considered in the evaluation of explanations.

In addition, the underlying algorithm of a recommender engine will influence the types of explanations that *can* be generated, even though the explanations selected by the system developer may or may not reflect the underlying algorithm. This is particularly the case for computationally complex algorithms for which explanations may be more difficult to generate, such as collaborative filtering (Herlocker et al., 2000; Hu et al., 2008). In this case, the developer must consider the trade-offs between e.g. satisfaction (as an extension of understandability) and transparency. In Section 3.8, we relate the most common explanation styles and how they relate to the underlying algorithms. Finally, we conclude with a summary in Section 3.9.

3.2 Guidelines

The content of this chapter is divided into sections which each elaborate on the following design guidelines for explanations in recommender systems.

- Consider the benefit(s) you would like to obtain from the explanations, and the best metric to evaluate on the associated criteria (Section 3.3).
- Be aware that the evaluation of explanations is related to, and may be confounded with, the functioning of the underlying recommendation engine, as measured by criteria commonly used for evaluating recommender systems (Section 3.5).
- Think about how the way that you present the recommendations themselves affects which types of explanations are more suitable (Section 3.6)
- Keep in mind how the interaction model that you select affects and interacts with the explanations (Section 3.7).
- Last, but certainly not least, consider how the underlying algorithm may affect the type of explanations you can generate (Section 3.8).

3.3 Defining criteria

Surveying the literature for explanations in recommender systems, we see that recommender systems with explanatory capabilities have been evaluated according to different criteria, and identify seven different criteria for explanations of single item recommendations. This compilation of explanatory criteria in recommender systems is the first of its kind. Table 3.1 states the criteria, which are similar to those desired (but not evaluated on) in expert systems, c.f. MYCIN (Bennett and Scott., 1985). In Table 3.2, we summarize previous evaluations of explanations in recommender systems, and the criteria by which they have been evaluated. Works that have no clear criteria stated, or have not evaluated the system on the explanation criteria which they state, are omitted from this table.

It is important to identify these criteria as distinct, even if they may interact, or require certain trade-offs. Indeed, it would be hard to create explanations that do well on all criteria, in reality it is a trade-off. For instance, in our work we have found that while personalized explanations may lead to greater user satisfaction, they do not necessarily increase effectiveness (Tintarev and Masthoff, 2008b) (see also Chapter 6). Other times, criteria that seem to be inherently related are not necessarily so, for example it has been found that transparency does not necessarily aid trust (Cramer et al., 2008b). For these reasons, while an explanation in Table 3.2 may have been evaluated for several criteria, it may not have achieved them all.

The type of explanation that is given to a user is likely to depend on the criteria of the designer of a recommender system. For instance, when building a system that sells books one might decide that user trust is the most important aspect, as it leads to user loyalty and increases sales. For selecting tv-shows, user satisfaction could be more important than effectiveness. That is, it is more important that a user enjoys the service, than that they are presented the best available shows.

In addition, some attributes of explanations may contribute toward achieving multiple goals. For instance, one can measure how *understandable* an explanation is, which can contribute to e.g. user trust, as well as satisfaction.

In this section we describe seven criteria for explanations, and suggest evaluation metrics based on previous evaluations of explanation facilities, or offer suggestions of how existing measures could be adapted to evaluate the explanation facility in a recommender system.

Table 3.1: Explanatory criteria

Aim	**Definition**
Transparency (Tra.)	Explain how the system works
Scrutability (Scr.)	Allow users to tell the system it is wrong
Trust	Increase users' confidence in the system
Effectiveness (Efk.)	Help users make good decisions
Persuasiveness (Pers.)	Convince users to try or buy
Efficiency (Efc.)	Help users make decisions faster
Satisfaction (Sat.)	Increase the ease of usability or enjoyment

3.3.1 Explain how the system works: Transparency

Transparency aims to help users understand how the system works, or in the case of explanations, understand how the recommendations were selected, or how the recommended item fits their needs. The importance of transparency is highlighted in an anecdotal article in the Wall Street Journal (Zaslow, 2002) titled "*If TiVo Thinks You Are Gay, Here's How*

Table 3.2: The criteria by which explanations in recommender systems have been evaluated. System names are mentioned if given, otherwise we only note the type of recommended items. Works that have no clear criteria stated, or have not *evaluated* the system on the explanation criteria which they state, are omitted from this table. Note that while a system may have been evaluated for several criteria, it may not have achieved all of them. Also, for the sake of completeness we have distinguished between multiple studies using the same system.

System (type of items)	**Tra.**	**Scr.**	**Trust**	**Efk.**	**Per.**	**Efc.**	**Sat.**
(Internet providers) (Felfernig and Gula, 2006)			X		X		X
(Digital cameras, notebooks computers) (Pu and Chen, 2006)			X				
(Digital cameras, notebooks computers) (Pu and Chen, 2007)			X	X			
(Music) (Sinha and Swearingen, 2002)			X				
(Movies) (Tintarev and Masthoff, 2008b)				X	X		X
Adaptive Place Advisor (restaurants) (Thompson et al., 2004)				X		X	
ACORN (movies) (Wärneståł, 2005b)							X
CHIP (cultural heritage artifacts) (Cramer et al., 2008a)	X		X	X			
CHIP (cultural heritage artifacts) (Cramer et al., 2008b)	X		X				X
iSuggest-Usability (music) (Hingston, 2006)	X			X			
LIBRA (books) (Bilgic and Mooney, 2005)				X			
MovieLens (movies) (Herlocker et al., 2000)					X		X
Moviexplain (movies) (Symeonidis et al., 2008)				X			X
myCameraAdvisor (Wang and Benbasat, 2007)			X				
Qwikshop (digital cameras) (McCarthy et al., 2004)				X		X	
SASY (e.g. holidays) (Czarkowski, 2006)	X	X					X
Tagsplanations (movies) (Vig et al., 2009)	X			X			

to Set It Straight" describes users' frustration with irrelevant choices made by a video recorder that records programs it assumes its owner will like, based on shows the viewer has recorded in the past. For example, one user, Mr. Iwanyk, suspected that his TiVo thought he was gay since it inexplicably kept recording programs with gay themes. This user clearly deserved an explanation.

An explanation may clarify *how* a recommendation was chosen. In expert systems, such as in the domain of medical decision making, the importance of transparency has long been recognized (Bennett and Scott., 1985). Transparency or the heuristic of "Visibility of System Status" is also an established usability principle (Nielsen and Molich, 1990), and its importance has also been highlighted in user studies of recommender systems (Sinha and Swearingen, 2002).

Vig et al. (2009) differentiate between transparency and justification. While transparency should give an honest account of how the recommendations are selected and how the system works, justification can be descriptive and decoupled from the recommendation algorithm. The authors cite several reasons for opting for justification rather than genuine transparency. For example some algorithms that are difficult to explain (e.g. latent semantic analysis where the distinguishing factors are latent and may not have a clear interpretation), protection of trade secrets by system designers, and the desire for greater freedom in designing the explanations.

Cramer et al. have investigated the effects of transparency on other evaluation criteria such as trust, persuasion (acceptance of items) and satisfaction (acceptance) in an art recommender (Cramer et al., 2008a,b). Transparency itself was evaluated in terms of its effect on actual and perceived understanding of how the system works (Cramer et al., 2008b). While actual understanding was based on user answers to interview questions, perceived understanding was extracted from self-reports in questionnaires and interviews.

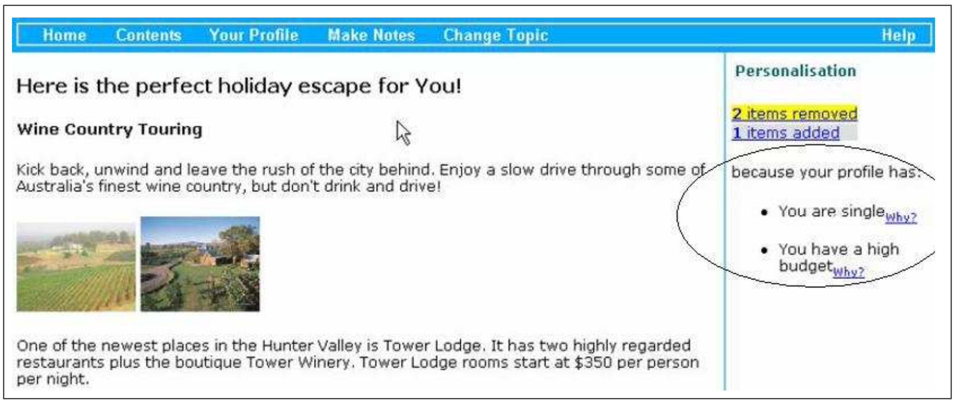
The evaluation of transparency has also been coupled with scrutability (Section 3.3.2) and trust (Section 3.3.3), but we will see in these sections that these criteria can be distinct from each other.

3.3.2 Allow users to tell the system it is wrong: Scrutability

Explanations may help isolate and correct misguided assumptions or steps in the process of recommendation. Explanations which help the user to change these assumptions or steps can be said to increase the *scrutability* of the recommender system. When a system collects and interprets information in the background, as is the case with TiVo, it becomes all the more important to make the reasoning available to the user. Following transparency, a second step is to allow a user to correct reasoning, or make the system *scrutable* (Czarkowski, 2006). Explanations should be part of a cycle, where the user understands what is going on in the system and exerts control over the type of recommendations made, by correcting system assumptions where needed (Sørmo et al., 2005). Scrutability is related to the established usability principle of User Control (Nielsen and Molich, 1990). See Figure 3.1 for an example of a scrutable holiday recommender.

While scrutability is very closely tied to the criteria of transparency, it deserves to be uniquely identified. There are explanations that are transparent, but not scrutable such as the explanation in Table 4.4. Here, the user cannot change (scrutinize) the ratings that affected this recommendation directly in the interface. In addition, the scrutability may reflect certain portions, rather the entire workings of the recommender engine. The explanations in this table are scrutable, but not (fully) transparent even if they offer some form of justification. For example, there is nothing in Table 4.4 that suggests that the underlying recommendations are based on a Bayesian classifier. In such a case, we can imagine that a user attempts to scrutinize a recommender system, and manages to change their recommendations, but still does not understand exactly what happens within the system.

Czarkowski found that users were not likely to scrutinize on their own, and that extra effort was needed to make the scrutability tool more visible (Czarkowski, 2006). In addition, it was easier to get users perform a given scrutinization task such as changing the personalization (e.g. *"Change the personalization so that only Current Affairs programs are included in your 4:30-5:30 schedule."*) Their evaluation included metrics such as task correctness, and if users could express an understanding of what information was used to make recommendations for them. They understood that adaptation in the system was based on their personal attributes stored in their profile, that their profile contained information they volunteered about themselves, and that they could change their profile to control the personalization (Czarkowski, 2006).



Here is the perfect holiday escape for You!

Wine Country Touring

Kick back, unwind and leave the rush of the city behind. Enjoy a slow drive through some of Australia's finest wine country, but don't drink and drive!

One of the newest places in the Hunter Valley is Tower Lodge. It has two highly regarded restaurants plus the boutique Tower Winery. Tower Lodge rooms start at $350 per person per night.

because your profile has:

- You are single Why?
- You have a high budget Why?

Figure 3.1: Scrutable adaptive hypertext, Czarkowski (2006). The explanation is in the circled area, and the user profile can be accessed via the "why" links.

3.3.3 Increase users' confidence in the system: Trust

We define trust as users' confidence in the system, e.g. that they believe that the recommendation process is correct and unbiased. Trust is sometimes linked with transparency: previous studies indicate that transparency and the possibility of interaction with recommender systems increases user trust (Felfernig and Gula, 2006; Sinha and Swearingen, 2002). Trust in the recommender system could also be dependent on the accuracy of the recommendation algorithm (McNee et al., 2003b). A study of users' trust (defined as perceived confidence in a recommender system's *competence*) suggests that users intend to return to recommender systems which they find trustworthy (Chen and Pu, 2002). We note however, that there is a study where transparency and trust were not found to be related (Cramer et al., 2008b).

We do not claim that explanations can fully compensate for poor recommendations, but good explanations may help users make better decisions (see Section 3.3.5 on effectiveness). A user may also be more forgiving, and more confident in recommendations, if they understand why a bad recommendation has been made and can prevent it from occurring again. A user may also appreciate when a system is "frank" and admits that it is not confident about a particular recommendation.

In addition, the design of the user interface for a recommender system may affect its credibility. In a study of factors determining web page credibility ("trustworthiness") the largest proportion of users' comments (46.1%) referred to the appeal of the overall visual design of a site, including layout, typography, font size and color schemes (Fogg et al., 2003). Likewise the perceived credibility of a Web article was significantly affected by the presence of a photograph of the author (Fogg et al., 2001). So while recommendation accuracy, and the criteria of transparency are often linked to the evaluation of trust, design is also a factor that needs to be considered as part of the evaluation.

Questionnaires can be used to determine the degree of trust a user places in a system. An overview of trust questionnaires can be found in Ohanian (1990) which also suggests and validates a five dimensional scale of trust. Note that this validation was done with the aim of using celebrities to endorse products, but was not conducted for a particular domain. Additional validation may be required to adapt this scale to a particular recommendation domain.

A model of trust in recommender systems is proposed in Chen and Pu (2002); Pu and Chen (2007), and the questionnaires in these studies consider factors such as intent to return to the system, and intent to save effort. Also Wang and Benbasat (2007) query users about trust, but focus on trust related beliefs such as the perceived competence, benevolence and integrity of a virtual adviser. Although questionnaires can be very focused, they suffer from the fact that self-reports may not be consistent with user behavior. In these cases, implicit measures (although less focused) may reveal factors that explicit measures do not.

One such implicit measure could be loyalty, a desirable bi-product of trust. One study compared different interfaces for eliciting user preferences in terms of how they affected factors such as loyalty (McNee et al., 2003b). Loyalty was measured in terms of the number of logins and interactions with the system. Among other things, the study found that allowing users to independently choose which items to rate affected user loyalty. It has also been thought that Amazon's conservative use of recommendations, mainly recommending familiar items, enhances user trust and has led to increased sales (Swearingen and Sinha, 2002).

3.3.4 Convince users to try or buy: Persuasiveness

Explanations may increase user evaluation of the given recommendations (Herlocker et al., 2000). We consider explanations that achieve this aim as persuasive, as they are an attempt to gain benefit for the system rather than for the user.

Cramer et al. (2008b) evaluated the acceptance of recommended items in terms of how many recommended items were present in a final selection of six favorites. In a study of a collaborative filtering- and rating-based recommender system for movies, participants were given different explanation interfaces (e.g. Figure 3.2) (Herlocker et al., 2000). This study directly inquired how likely users were to see a movie (with identifying features such as title omitted) for 21 different explanation interfaces. Persuasion was thus a numerical rating on a 7-point Likert scale.

Persuasiveness can be measured in a number of ways, for example, it can be measured as the difference between two ratings: the first being a previous rating, and the second a re-rating for the same item, but with an explanation interface (Cosley et al., 2003). Indeed, it has been shown that users can be manipulated to give a rating closer to the system's prediction (Cosley et al., 2003). This study was in the low investment domain of movie rental, and it is possible that users may be less influenced by incorrect predictions in high(er) cost domains such as cameras[1]. It is also important to consider that too much persuasion may backfire once users realize that they have tried or bought items that they do not really want.

Another possibility would be to measure how much users actually try or buy items compared to users in a system without an explanation facility. These metrics can also be understood in terms of the concept of "conversion rate" commonly used in e-Commerce, operationally defined as the percentage of visitors who take a desired action.

[1] In Tintarev and Masthoff (2008a) participants reported that they found incorrect overestimation less useful in high cost domains compared to low cost domains (see also Chapter 4 and Section 4.5 in particular)

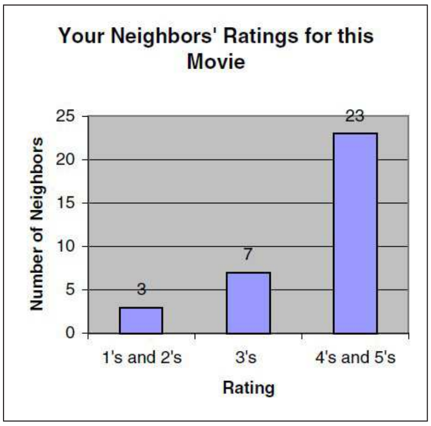


Figure 3.2: One out of twenty-one interfaces evaluated for persuasiveness - a histogram summarizing the ratings of similar users (neighbors) for the recommended item grouped by good (5 and 4's), neutral (3s), and bad (2s and 1s), on a scale from 1 to 5 Herlocker et al. (2000).

3.3.5 Help users make good decisions: Effectiveness

Rather than simply persuading users to try or buy an item, an explanation may also assist users to make *better* decisions, or be effective. Effectiveness is by definition highly dependent on the accuracy of the recommendation algorithm, i.e. users cannot make correct decisions if the recommendations themselves are faulty. An effective explanation would help the user evaluate the quality of suggested items according to their own preferences. This would increase the likelihood that the user discards irrelevant options while helping them to recognize useful ones. For example, a book recommender system with effective explanations would help a user to buy books they actually end up liking. Bilgic and Mooney (2005) emphasize the importance of measuring the ability of a system to assist the user in making accurate decisions about recommendations, and compared different explanation types for effectiveness. Effective explanations could also serve the purpose of introducing a new domain, or the range of products, to a novice user, thereby helping them to understand the full range of options (Felfernig and Gula, 2006; Pu and Chen, 2006).

Vig et al. measure perceived effectiveness: *"This explanation helps me determine how well I will like this movie."* (Vig et al., 2009). Effectiveness of explanations can also be calculated as the *absence of a difference* between the liking of the recommended item prior to, and after, consumption. For example, in a previous study, users rated a book twice, once after receiving an explanation, and a second time after reading the book (Bilgic and Mooney, 2005). If their opinion on the book did not change much, the system was considered effective. This study explored the effect of the whole recommendation process, explanation inclusive, on effectiveness. We have used the same metric to evaluate whether personalization of explanations (in isolation of a recommender system) increased their effectiveness in the movie domain (Tintarev and Masthoff, 2008a) (see also Chapter 6).

While this metric considers the difference between the before and after ratings, it does not discuss the effects of over- contra underestimation. In our work we found that users considered overestimation to be less effective than underestimation, and that this varied between domains. Specifically, overestimation was considered more severely in high investment domains compared to low investment domains. In addition, the strength of the effect on *perceived* effectiveness varied depending on where on the scale the prediction error occurred (Tintarev and Masthoff, 2008a) (see also Chapter 4 and Section 4.5 in particular).

Another way of measuring the effectiveness of explanations has been to test the same system with and without an explanation facility, and evaluate if subjects who receive explanations end up with items more suited to their personal tastes (Cramer et al., 2008a).

Other work evaluated explanation effectiveness using a metric from marketing (Häubl and Trifts, 2000), with the aim of finding the single *best* possible item (rather than “good enough items” as above) (Chen and Pu, 2007). Participants interacted with the system until they found the item they would buy. They were then given the opportunity to survey the entire catalog and to change their choice of item. Effectiveness was then measured by the fraction of participants who found a better item when comparing with the complete selection of alternatives in the database. So, using this metric, a low fraction represents high effectiveness.

Effectiveness is the criterion that is most closely related to accuracy measures such as precision and recall (Cramer et al., 2008a; Symeonidis et al., 2008; Thompson et al., 2004). In systems where items are easily consumed, e.g. internet news, these can be translated into recognizing relevant items and discarding irrelevant items respectively. For example, there have been suggestions for an alternative metric of “precision” based on the number of profile concepts matching with user interests, divided by the number of concepts in their profile (Cramer et al., 2008a).

We have chosen to focus this thesis primarily on this criterion. While many of the other criteria had much scope left for exploration, at the time this thesis began there was particularly little empirical data on what constitutes an effective explanation. In addition, we remind the reader that while we consider effectiveness as the main criteria, we do consider other criteria as well. We will elaborate the criterion of effectiveness in greater detail in Chapter 4.

3.3.6 Help users make decisions faster: Efficiency

Efficient explanations would make it *faster* for users to decide which recommended item is best for them. Efficiency is another established usability principle, i.e. how quickly a task can be performed (Nielsen and Molich, 1990). This criterion is often addressed in the recommender systems literature (See Table 3.2) given that the task of recommender systems is to find needles in haystacks of information.

Efficiency may be improved by allowing the user to understand the relation between competing options (McCarthy et al., 2004; McSherry, 2005; Pu and Chen, 2006). In the domain of digital cameras, competing options may for example be described as ”Less Memory and Lower Resolution and Cheaper” (McCarthy et al., 2004). This way users are *quickly* able to find something cheaper if they are willing to settle for less memory and lower resolution, and do not need to keep searching for something better.

Efficiency is often used in the evaluation of so-called conversational recommender systems, where users continually interact with a recommender system, refining their preferences (see also Section 3.7.1). In these systems, the explanations can be seen to be implicit in the dialog. Efficiency in these systems can be measured by the total amount of interaction time, and number of interactions needed to find a satisfactory item (Thompson et al., 2004). Evaluations of explanations based on improvements in efficiency are not limited to conversational systems however. Pu and Chen for example, compared completion time for two explanatory interfaces, and measured completion time as the amount of time it took a participant to locate a desired product in the interface (Pu and Chen, 2006).

Other metrics for efficiency also include the number of inspected explanations, and number of activations of repair actions when no satisfactory items are found (Felfernig and Gula, 2006; Reilly et al., 2004a). Normally, it is not sensible to expose users to all possible recommendations and their explanations, and so users can choose to inspect (or scrutinize) a given recommendation by asking for an explanation. In a more efficient system, the users would need to inspect *fewer* explanations. Repair actions consist of feedback from the user which changes the type of recommendation they receive, as outlined in the sections on scrutability (Section 3.3.2). Examples of user feedback/repair actions can be found in Section 3.7.

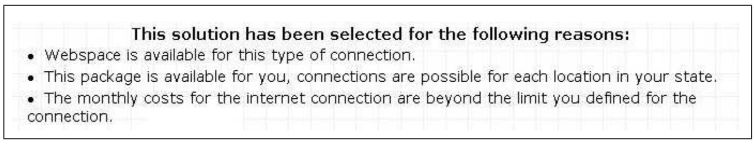
3.3.7 Make the use of the system fun: Satisfaction

Explanations such as those in (Tanaka-Ishii and Frank, 2000), aim to increase the acceptance of the system as a whole, or increase satisfaction. The presence of longer descriptions of individual items has been found to be positively correlated with both the *perceived* usefulness and ease of use of the recommender system (Sinha and Swearingen, 2002). This can be seen as improving users’ overall satisfaction. Also, many commercial recommender systems such as those seen in Table 3.4 are primarily sources of entertainment. In these cases, any extra facility should take notice of the effect on user satisfaction. Figure 3.3 gives an example of an explanation which the authors claim is aimed at increasing satisfaction.

When measuring satisfaction, one can directly ask users whether the system is enjoyable to use. Tanaka-Ishii and Frank in their evaluation of a multi-agent system describing a Robocup soccer game ask users whether they prefer the system with or without explanations (Tanaka-Ishii and Frank, 2000). Satisfaction can also be measured indirectly by measuring user loyalty (McNee et al., 2003b; Felfernig and Gula, 2006) (see also Section 3.3.3), and likelihood of using the system for a search task (Cramer et al., 2008b).

In measuring explanation satisfaction, it is important to differentiate between satisfaction with the recommendation process[2], and the recommended products (persuasion) (Cramer et al., 2008b; Felfernig and Gula, 2006). One (qualitative) way to measure satisfaction with the process would be to conduct a user walk-through for a task such as finding a satisfactory item. In such a case, it is possible to identify usability issues and even apply quantitative metrics such as the ratio of positive to negative comments; the number of times the evaluator was frustrated; the number of times the evaluator was delighted; the number of times and where the evaluator worked around a usability problem etc.

It is also arguable that users would be satisfied with a system that offers effective explanations, confounding the two criteria. However, a system that aids users in making good decisions, may have other disadvantages that decrease the overall satisfaction (e.g. requiring a large cognitive effort on the part of the user). Fortunately, these two criteria can be measured by distinct metrics.



This solution has been selected for the following reasons:

- Webspace is available for this type of connection.
- This package is available for you, connections are possible for each location in your state.
- The monthly costs for the internet connection are beyond the limit you defined for the connection.

Figure 3.3: An explanations for an internet provider, describing the provider in terms of user requirements: "This solution has been selected for the following reasons ..." Felfernig and Gula (2006)

3.4 Explanations in expert systems

Explanations in intelligent systems are not a new idea: explanations have often been considered as part of the research in the area of expert systems (Andersen et al., 1990; Hunt and Price, 1988; Lopez-Suarez and Kamel, 1994; Hance and Buchanan, 1984; Wick and Thompson, 1992). Both types of system can be used to support a decision-making process, but there are also some fundamental differences. For example, while recommender

[2] Here we mean the entire recommendation process, inclusive of the explanations. However, in Section 3.5 we highlight that evaluation of explanations in recommender systems are seldom fully independent of the underlying recommendation process.

systems consider many simpler cases, expert systems are often used to make a complex decision about a single problem. The research on explanations in expert systems has largely been focused on what kind of explanations can be generated and how these have been implemented in real world systems (Andersen et al., 1990; Hunt and Price, 1988; Lopez-Suarez and Kamel, 1994; Wick and Thompson, 1992). Overall, *there are few evaluations of the explanations in these systems* (see also Section 3.4.4).

Also, developments in recommender systems have revived explanation research, after a decline of studies in expert systems in the 90's. One such development is the increase in data: due to the growth of the web, there are now more users using the average (recommender) system. Systems are also no longer developed in isolation of each other, making the best possible reuse of code (open source projects) and datasets (e.g. the MovieLens[3] and Netflix datasets[4]). In addition, new algorithms have been adapted and developed (e.g. kNN (Breese et al., 1998; Resnick et al., 1994), latent semantic analysis (Deerwester et al., 1990; Hofmann, 2003), which mitigate domain dependence, and allow for greater generalizability. One sign of the revived interest in explanation research is the success of a recent series of workshops on explanation aware computing (Roth-Berghofer et al., 2008, 2009).

Expert systems can roughly be split into three families of methods: heuristic-based methods (rule-based), Bayesian networks, and case-based reasoning (commonly abbreviated as CBR). Reviews of expert systems with explanatory capabilities in each family can be found in (Lacave and Diéz, 2004) (heuristic), (Lacave and Diéz, 2002) (bayesian) and (Doyle et al., 2003) (CBR) respectively. Please note that this is not meant as a comprehensive review, but as a brief summary relating how this body of work ties into our own.

3.4.1 Heuristic-based

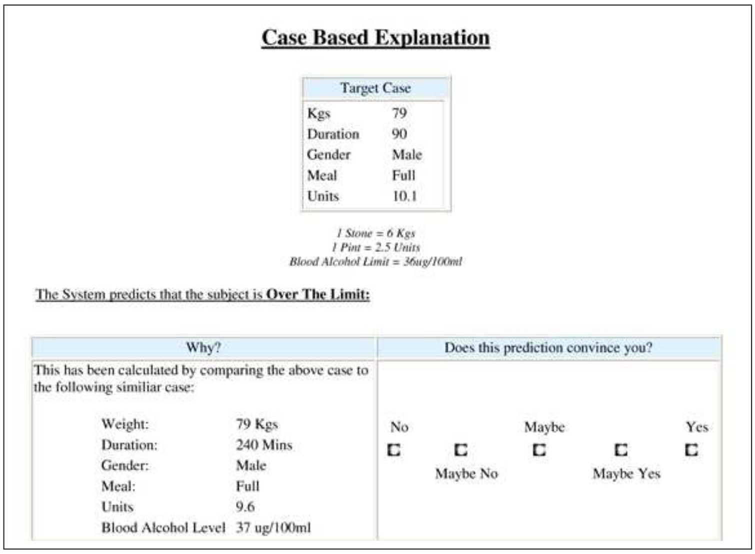
Heuristic-based expert systems use rules such as *"If blood sugar > 100 + 3.7 * (150-Na) = Yes GOTO Questions Concerning Hypernatremia"*, often chaining them together into a line of reasoning. The chain of used rules then serves as a type of explanation, such as in Figure 3.5. This family of systems is the most adaptive and considers the users' level of knowledge (Lacave and Diéz, 2004), such as considering the amount of detail to use in the explanation (Bleich, 1972), the users' information needs, or what has been mentioned by the system previously (Carenini et al., 1994).

[3] http://www.grouplens.org/node/73, retrieved July 2009

[4] http://www.netflixprize.com/, retrieved July 2009

3.4.2 Bayesian networks

Explanations in expert systems using Bayesian networks are based on probabilities, and can be hard to express in a user-understandable manner. Druzdzel (1996) has looked at ways of expressing probabilities in plain English, in which causal relations play an important role. That is, one needs to know which factors affect each other, and which is the source and which is the cause such as in: *"Cold very commonly (p=0.9) causes sneezing. Allergy very commonly (p=0.9) causes sneezing. Cold does not affect the tendency of allergy to cause sneezing, and vice versa"*. Interaction with users in these systems has been highly limited, and they do not adapt to users (Lacave and Diéz, 2002).



Case Based Explanation

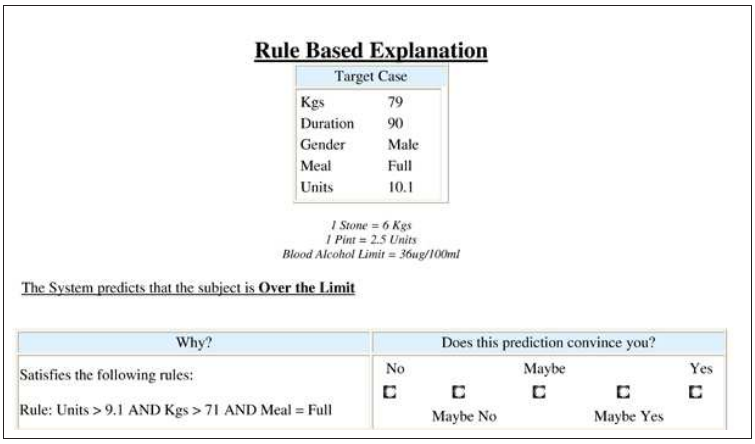
Target Case	
Kgs	79
Duration	90
Gender	Male
Meal	Full
Units	10.1

1 Stone = 6 Kgs
1 Pint = 2.5 Units
Blood Alcohol Limit = 36ug/100ml

The System predicts that the subject is **Over The Limit:**

Why?	Does this prediction convince you?
This has been calculated by comparing the above case to the following similiar case: Weight: 79 Kgs Duration: 240 Mins Gender: Male Meal: Full Units 9.6 Blood Alcohol Level 37 ug/100ml	No □ □ Maybe No □ Maybe □ Maybe Yes □ Yes

Figure 3.4: Example of a case-based prediction and explanation (Cunningham et al., 2003)



Rule Based Explanation

Target Case	
Kgs	79
Duration	90
Gender	Male
Meal	Full
Units	10.1

1 Stone = 6 Kgs
1 Pint = 2.5 Units
Blood Alcohol Limit = 36ug/100ml

The System predicts that the subject is **Over the Limit**

Why?	Does this prediction convince you?
Satisfies the following rules: Rule: Units > 9.1 AND Kgs > 71 AND Meal = Full	No □ □ Maybe No □ Maybe □ Maybe Yes □ Yes

Figure 3.5: Example of a heuristic-based prediction and explanation (Cunningham et al., 2003)

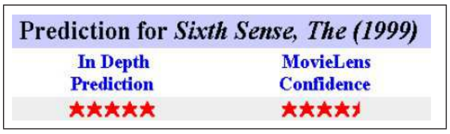
3.4.3 Case-based Reasoning (CBR)

CBR expert systems use a knowledge-base containing examples of previous cases, e.g. medical files, that are similar to the current problem, and use them to predict a solution to the current problem. Formally, the CBR method is defined in terms of these three tasks: 1) retrieve, that obtains past cases similar to the new case; 2) select, that decides which of the retrieved past cases is the most similar (i.e. the best precedent) to the current problem; and 3) adapt, that decides how to adapt the solution of the best precedent to solve the current problem (Armengol and Plaza, 1994). Doyle et al. (2003) summarizes some user adaptations in CBR expert systems such as those by (Finch, 1998) and (Wolverton, 1995). The explanations in these systems can adapt to the intended recipient (Finch, 1998), or the users' knowledge and beliefs (Wolverton, 1995).

3.4.4 Evaluations in expert systems

Research on explanation facilities in expert systems has largely been focused on what kind of explanations can be generated and how these have been implemented in real world systems (Andersen et al., 1990; Hunt and Price, 1988; Lopez-Suarez and Kamel, 1994; Wick and Thompson, 1992). Evaluations of these systems have largely focused on user acceptance such as (Carenini et al., 1994), and in some cases the decision support of the system as a whole has been evaluated (Hance and Buchanan, 1984). User acceptance can be defined in terms of our criteria of satisfaction or persuasion. If the evaluation measures acceptance with the system as whole, such as Carenini et al. (1994) who asked questions such as *"Did you like the program?"*, this reflects user satisfaction. If rather, the evaluation measures user acceptance of advice or explanations, (e.g. Ye et al., 1995), the aim can be said to be persuasion.

In recent user studies, there are also some particularly notable works. Cunningham et al. (2003) compared case-based explanations (see Figure 3.4) with heuristic-based explanations (see Figure 3.5), and found the case-based explanations to be more persuasive. In fact, rule-based explanations did not perform noticeably better than no explanation at all. Ye et al. (1995) studied the role of explanations on user acceptance. They found that explanations could increase user acceptance of the system's conclusions, and that justifying explanations using reasoning and causal arguments were particularly persuasive. The design used in Ye et al. (1995) is similar to that of Cosley et al. (2003): they both compared the belief in advice/items before and after explanations (ratings in the case of Cosley et al. (2003)), to see if they increased persuasion.



Prediction for *Sixth Sense, The (1999)*

In Depth Prediction	MovieLens Confidence
★★★★★	★★★★½

Figure 3.6: Confidence display for a recommendation, Herlocker et al. (2000) - the movie is strongly recommended (5/5), and there is a large amount of information to support the recommendation (4.5/5).

3.5 Evaluating the impact of explanations on the recommender system

We have now identified seven criteria by which explanations in recommender systems can be evaluated, and given suggestions of how such evaluations can be performed. To some extent, these criteria assume that we are evaluating only the explanation component. It also seems reasonable to evaluate the system as a *whole*. In that case we might measure the general system usability and accuracy, which will depend on both the recommendation algorithm as well as the impact of the explanation component. Therefore, in this section, we describe the interaction between the recommender engine and our explanation criteria, organized by the evaluation metrics commonly used in recommender system evaluations: accuracy, learning rate, coverage, novelty/serendipity and acceptance.

3.5.1 Accuracy Metrics

Accuracy metrics regard the ability of the recommendation engine to predict correctly, but accuracy is likely to interact with explanations too. For example, the relationship between transparency and accuracy is not self-evident: Cramer et al. found that transparency led to changes in user behavior that ultimately decreased recommendation accuracy (Cramer et al., 2008a).

The system's own confidence in its recommendations is also related to accuracy and can be reflected in explanations. An example of an explanation aimed to help users understand (lack of) accuracy, can be found in confidence displays such as Figure 3.6. These can be used to explain e.g. poor recommendations in terms of insufficient information used for forming the recommendation. For further work on confidence displays see also (McNee et al., 2003a).

Explanations can also help users understand how they would relate to a particular item, possibly supplying additional information that helps the user make more informed decisions (effectiveness). In the case of poor accuracy, the risk of missing good items, or trying bad ones increases while explanations can help decrease this risk. By helping users to correctly identify items as good or bad, the accuracy of the recommender system as a whole may also increase.

3.5.2 Learning Rate

The learning rate represents how quickly a recommender system learns a user's preferences, and how sensitive it is to changes in preferences. Learning rate is likely to affect user satisfaction as users would like a recommender system to quickly learn their preferences, and be sensitive to short term as well as long term interests. Explanations can increase satisfaction by clarifying or hinting that the system considers changes in the user's preferences. For example, the system can flag that the value for a given variable is getting close to its threshold for incurring a change, but that it has not reached it yet. A system can also go a step further, and allow the user to see just how it is learning and changing preferences (transparency), or make it possible for a user to delete old preferences (scrutability). For example, the explanation facility can request information that would help it learn/change quicker, such as asking if a user's favorite movie genre has changed from action to comedy.

3.5.3 Coverage

Coverage regards the range of items which the recommender system is able to recommend. Explanations can help users understand where they are in the search space. By directing the user to rate informative items in under-explored parts of the search space, explanations may increase the overlap between certain items or features (compared to sparsity). Ultimately, this may increase the overall coverage for potential recommendations. Understanding the remaining search options is related to the criterion of transparency: a recommender system can explain why certain items are not recommended. It may be impossible or difficult to retrieve an item (e.g. for items that have a very particular set of properties in a knowledge-based system, or the item does not have many ratings in a collaborative-filtering system). Alternatively, the recommender system may function under the assumption that the user is not interested in the item (e.g. if their requirements are too narrow in a knowledge-based system, or if they belong to a very small niche in a collaborative-based system). An explanation can explain why an item is not available for recommendation, and even how to remedy this and allow the user to change their preferences (scrutability).

Coverage may also affect evaluations of the explanatory criteria of effectiveness. For example, if a user's task is not only to find a "good enough" item, but the best item for them, the coverage needs to be sufficient to ensure that "best" items are included in the recommendations. Depending on how much time retrieving these items takes, coverage may also affect efficiency.

3.5.4 Acceptance

It is possible to confound acceptance, or satisfaction with a system with other types of satisfaction. If users are satisfied with a system with an explanation component, it remains unclear whether this is due to: satisfaction with the *explanation component*, satisfaction with *recommendations*, or general design and visual appeal. Satisfaction with the system due to the recommendations is connected to accuracy metrics, or even novelty and diversity, in the sense that sufficiently good recommendations need to be given to a user in order to keep them satisfied. Although explanations may help increase satisfaction, or tolerance toward the system, they cannot function as a substitute for e.g. good accuracy. Indeed, this is true for all the mentioned explanatory criteria. An example of an explanation striving toward the criterion of satisfaction may be: *"Please bear with me, I still need to learn more about your preferences before I can make an accurate recommendation."*

3.6 Presenting recommendations

Some ways of presenting recommendations affect the explanation more than others. In fact, some ways of offering recommendations, such as the organizational structure we will describe shortly (see Section 3.6.5), can be seen as an explanation in itself. This (the explanation + presentation) in turn has an effect on the explanation criteria. The effect of presentational choices on the different explanation criteria is still an area with many under-explored aspects however.

The way a recommendation is presented may also show how good or relevant the item is considered to be. Relevance can be represented by the order in which recommendations are given in a list, e.g. with the best items at the top. When a single item is recommended, it tends to be the best one available. Relevance can also be visualized using e.g. different colors and font sizes, or shown via ratings. Ratings can use different scales and different symbols such as numbers or stars. Below we mention ways of offering recommendations in more detail, and illustrate how explanations may be used in each case.

3.6.1 Top item

Perhaps the simplest way to present a recommendation is by offering the user the best item. The way in which this item is selected could then be used as part of the explanation. Let us imagine a user who is interested in sport items, and appreciates football, but not tennis or hockey. The recommender system could then offer a recent football item, for example regarding the final in the world cup. The generated explanation may then be the following: *"You have been watching a lot of sports, and football in particular. This is the most popular and recent item from the world cup."* Note that this example uses the user's viewing history in order to form the explanation. A system could also use information that the user specifies more directly, e.g. how much they like football.

3.6.2 Top N-items

The system may also present several items at once. In a large domain such as news, it is likely that a user has many interests. In this case there are several items that could be highly interesting to the user. If the football fan mentioned above is also interested in technology news, the system might present several sports stories alongside a couple of technology items. Thus, the explanation generated by the system might be along the lines of: *"You have watched a lot of football and technology items. You might like to see the local football results and the gadget of the day."* The system may also present several items on the same theme, such as several football results. Note that while this system should be able to explain the relation between chosen items, it should still be able to explain the rational behind each single item.

3.6.3 Similar to top item(s)

Once a user shows a preference for one or more items, the recommender system can offer *similar* items. For each item, it can present one or more similar items (e.g. a list), and may show explanations similar to the ones in Sections 3.6.1 and 3.6.2. For example, given that a user liked a book by *Charles Dickens* such as Great Expectations, the system may present a recommendation together with this previously liked item in the following way; *"You might also like...Oliver Twist by Charles Dickens"*.

A recommender system can also offer recommendations in a social context, taking into account users that are similar to you. For example a recommendation can be presented in the following manner; *"People like you liked..Oliver Twist by Charles Dickens"*.

3.6.4 Predicted ratings for all items

Rather than forcing selections on the user, a system may allow its users to browse all the available options. Recommendations are then presented as predicted ratings on a scale (say from 0 to 5) for each item. A user may then still find items with low *predicted* ratings, and can counteract predictions by rating the affected items, or directly modifying the user model, i.e. changing the system's view of their preferences. This allows the user to tell the system when it is wrong, fulfilling the criteria of scrutability (see Section 3.3.2). Let us re-use our example of the football and technology fan. This type of system might on average offer higher predictions for football items than hockey items. A user might then ask why a certain item, for example local hockey results, is predicted to have a low rating. The recommender system might then generate an explanation like: *”This is a sports item, but it is about hockey. You do not seem to like hockey!”*.

If the user is interested in local hockey results, but not in results from other countries, they might modify their user model to limit their interest in hockey to local sports.

3.6.5 Structured overview

Pu and Chen (2006) suggest a structure which displays trade-offs between items. The best matching item is displayed at the top. Below it, several categories of trade-off alternatives are listed. Each category has a title explaining the characteristics of the items in it, e.g. *“[these laptops]...are cheaper and lighter, but have lower processor speed”*. The order of the titles depends on how well the category matches the user’s requirements.

Yee et al. (2003) used a multi-faceted approach for museum search and browsing. This approach considers *several* aspects of each item, such as location, date and material, each with a number of levels. The user can see how many items there are available at each level for each aspect. Using multiple aspects might be a suitable approach for a large domain with many varying objects.

Although not yet used in recommender systems, the ”treemap” structure (see Figure 3.7 [5]) allows a different type of overview (Bederson et al., 2002). Here it is possible to use different colors to represent topic areas, square and font size to represent importance to the current user, and shades of each topic color to represent recency.

The advantage of a structured overview is that the user can see ”where” they are in the search space, and possibly how many items can be found in each category. This greatly facilitates both navigation and user comprehension of the available options.

3.6.6 Recommender “Personality”

The choice of recommended items, or the predicted rating for an item can be angled to reflect a “personality” of the recommender system (McNee et al., 2006a). The recommender may have an *affirming* personality, supplying the user with recommendations of items they might already know about. This could inspire a user’s *trust* (see Section 3.3.3) in the system’s ability to present relevant or accurate items. Or, on the contrary, it may aim to offer more *novel* and positively surprising (serendipitous) recommendations in order to increase user *satisfaction*.

When a recommendation is made, it is operating along two often conflicting dimensions (Herlocker et al., 2004). The first dimension is the strength of the recommendation: how much does the recommender system think the user will like this item. The second dimension is the confidence of the recommendation: how sure is the recommender system

[5] http://www.marumushi.com/apps/newsmap/index.cfm, retrieved Aug. 2008

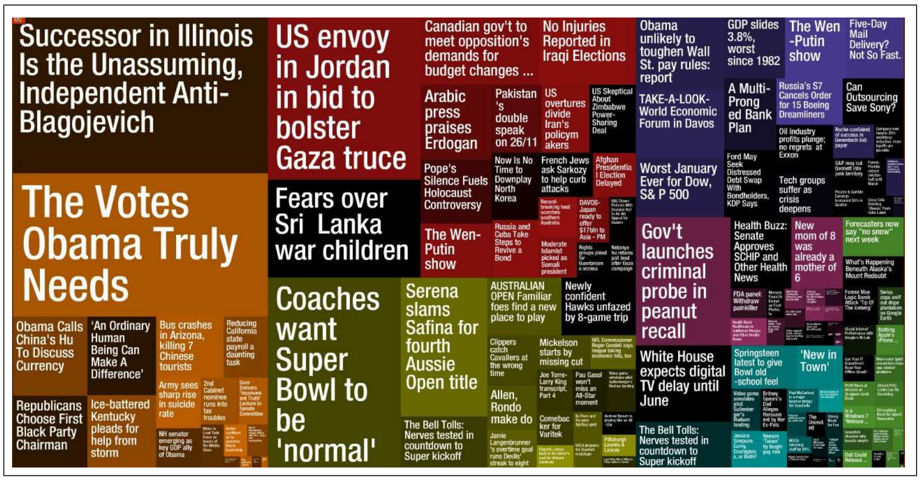

Successor in Illinois Is the Unassuming, Independent Anti-Blagojevich
US envoy in Jordan in bid to bolster Gaza truce
Canadian gov't to meet opposition's demands for budget changes ...
No Injuries Reported in Iraqi Elections
Obama unlikely to toughen Wall St. pay rules: report
GDP slides 3.8%, worst since 1982
The Wen -Putin show
Five-Day Mail Delivery? Not So Fast.
Arabic press praises Erdogan
Pakistan 's double speak on 26/11
US overtures divide Iran's policymakers
TAKE-A-LOOK- World Economic Forum in Davos
A Multi-Prong ed Bank Plan
Russia's S7 Cancels Order for 15 Boeing Dreamliners
Can Outsourcing Save Sony?
The Votes Obama Truly Needs
Fears over Sri Lanka war children
Pope's Silence Fuels Holocaust Controversy
Worst January Ever for Dow, S&P 500
Tech groups suffer as crisis deepens
The Wen-Putin show
Gov't launches criminal probe in peanut recall
Health Buzz: Senate Approves SCHIP and Other Health News
New mom of 8 was already a mother of 6
Forecasters now say "no snow" next week
Coaches want Super Bowl to be 'normal'
Serena slams Safina for fourth Aussie Open title
AUSTRALIAN OPEN Familiar foes find a new place to play
Newly confident Hawks unfazed by 8-game trip
Obama Calls China's Hu To Discuss Currency
'An Ordinary Human Being Can Make A Difference'
Bus crashes in Arizona, killing 7 Chinese tourists
Republicans Choose First Black Party Chairman
Ice-battered Kentucky pleads for help from storm
White House expects digital TV delay until June
Springsteen latest to give Bowl old -school feel
'New in Town'
Allen, Rondo make do
The Bell Tolls: Nerves tested in countdown to Super kickoff

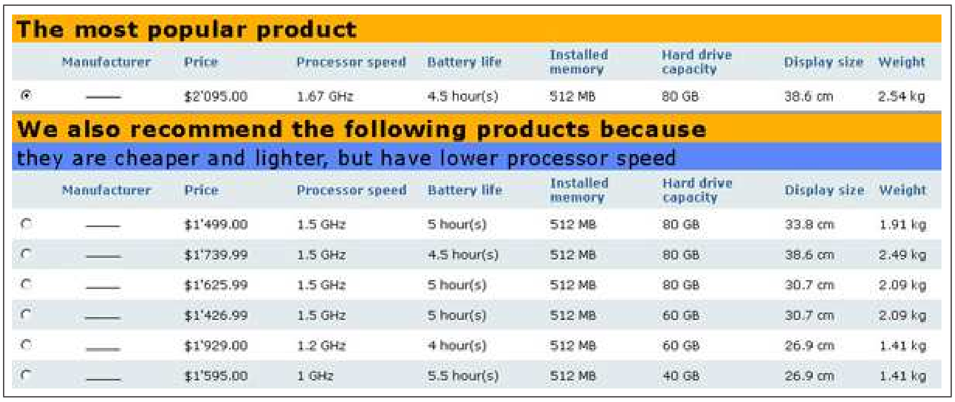
Figure 3.7: Newsmap - a treemap visualization of news. Different colors represent topic areas, square and font size to represent importance to the current user, and shades of each topic color to represent recency.

that its recommendation is accurate. A recommender system can be *bold* and recommend items more strongly than it normally would, or it could simply state its true *confidence* in its own recommendation (Herlocker et al., 2000).

If such factors are part of the recommendation process, the criteria of transparency (see Section 3.3.1) suggests that they should be part of the explanations as well.

3.7 Interacting with the recommender system

This section is dedicated to different ways in which a user can interact with a recommender system to influence the recommendations that they are given. This type of interaction is what distinguishes conversational systems from "single-shot" recommendations.



The most popular product

	Manufacturer	Price	Processor speed	Battery life	Installed memory	Hard drive capacity	Display size	Weight
●	——	$2'095.00	1.67 GHz	4.5 hour(s)	512 MB	80 GB	38.6 cm	2.54 kg

We also recommend the following products because
they are cheaper and lighter, but have lower processor speed

	Manufacturer	Price	Processor speed	Battery life	Installed memory	Hard drive capacity	Display size	Weight
○	——	$1'499.00	1.5 GHz	5 hour(s)	512 MB	80 GB	33.8 cm	1.91 kg
○	——	$1'739.99	1.5 GHz	4.5 hour(s)	512 MB	80 GB	38.6 cm	2.49 kg
○	——	$1'625.99	1.5 GHz	5 hour(s)	512 MB	80 GB	30.7 cm	2.09 kg
○	——	$1'426.99	1.5 GHz	5 hour(s)	512 MB	60 GB	30.7 cm	2.09 kg
○	——	$1'929.00	1.2 GHz	4 hour(s)	512 MB	60 GB	26.9 cm	1.41 kg
○	——	$1'595.00	1 GHz	5.5 hour(s)	512 MB	40 GB	26.9 cm	1.41 kg

Figure 3.8: Organizational Structure, Pu and Chen (2006)

They allow users to elaborate their requirements over the course of an extended dialog (Rafter and Smyth, 2005) rather than each user interaction being treated independently of previous history.

While interaction can occur in a recommender system regardless of the explanatory aim, it may be arguable that this type of interaction is particularly relevant to the explanation aims of transparency, effectiveness, and scrutability. Justifying recommendations, or making them transparent, to the user is only half of the solution, the second half is making the system *scrutable* (see Section 3.3.2) by allowing the user to make changes. By allowing users to change their preferences the explanation facility may also increase effectiveness when this leads to more relevant recommendations.

There are different ways in which a user can give feedback to the system to let it know how it is doing. Here we expand on four ways suggested by (Ginty and Smyth, 2002), supplying examples of current applications. Note that although unobtrusive elicitation of user preferences, via e.g. usage data (O'Sullivan et al., 2004) or demographics (Adrissono et al., 2003) can sometimes be more effective (O'Sullivan et al., 2004), this section focuses on *explicit* feedback from users.

3.7.1 The user specifies their requirements

By allowing a user to directly specify their requirements it is possible to circumvent the type of faulty assumptions that can be made by a system where the interests of a user are based on the items they decide to see, or how they interact with them. An indirect approach depends on the way it collects data, and is sometimes hard to change in the intended direction. Remember Mr. Iwanyk? This user tried to tame TiVo's gay fixation by recording war movies and other "guy stuff". As a result it started to give him documentaries on Joseph Goebbels and Adolf Eichmann. It stopped thinking he was gay and decided he was a crazy guy reminiscing about the Third Reich.

One movie recommender system allows users to specify their preferences in plain English (Wärnestål, 2005a), p. 77:

"**User:** I feel like watching a thriller.
System: Can you tell me one of your favorite thriller movies.
User: Uhm, I'm not sure.
System: Okay. Can you tell me one of your favorite actors or actresses?
User: I think *Bruce Willis* is good.
System: I see. Have you seen *Pulp Fiction*?
User: No.
System: Pulp Fiction is a thriller starring *Bruce Willis*. "

This dialog does not make use of the user's previous interests, nor does it explain *directly*. That is, there is no sentence that claims to be a justification of the recommendation.

It does however do so indirectly, by reiterating (and satisfying) the user's *requirements*. The user should then be able to interact with the recommender system, and give their opinion of the recommendation, thus allowing further refinement.

3.7.2 The user asks for an alteration

A more direct approach is to allow users to explicitly ask for alterations to recommended items, for instance using a structured overview (see also Section 3.6.5), or directly making the equivalent alteration McCarthy et al. (2004). This approach helps the users to find what they want quicker. Users can see how items compare, and see what other items are still available if the current recommendation should not meet their requirements. Have you ever put in a search for a flight, and been told to try other dates, other airports or destinations? This answer does not explain which of your criteria needs to be changed, requiring you to go through a tiring trial-and-error process. If you can see the trade-offs between alternatives from the start, and make the necessary alternations to your search, the initial problem can be circumvented.

Some feedback facilities allow users to see how criteria affect their remaining options. One such system explains the difference between a selected camera and remaining cameras. For example, it describes competing cameras with "Less Memory and Lower Resolution and Cheaper" (McCarthy et al., 2004). The user can ask for a more detailed explanation of the alternative criteria, and have a look at the cameras which fulfill these criteria. Instead of simply explaining to a user that no items fitting the description exist, these systems show what types of items *do* exist. These methods have the advantage of helping users find good enough items, even if some of their initial requirements were too strict.

3.7.3 The user rates items

To change the type of recommendations they receive, the user may want to correct predicted ratings, or modify a rating they made in the past. Ratings may be explicitly inputted by the user, or inferred from usage patterns. In a book recommender system a user could see the influence (in percentage) their previous ratings had on a given recommendation (Bilgic and Mooney, 2005). The *influence based explanation* showed which rated titles influenced the recommended book the most (see Figure 3.3). Although this particular system did not allow the user to modify previous ratings, or degree of influence, in the explanation interface, it can be imagined that this functionality could be implemented. Note however, that ratings may be easier to modify than the degree of influence which is likely to be computed.

Table 3.3: Influence of previous book ratings, on the current book recommendation (Bilgic and Mooney, 2005)

BOOK	**YOUR RATING** **Out of 5**	**INFLUENCE** **Out of 100**
Of Mice and Men	4	54
1984	4	50
Till We Have Faces: A Myth Retold	5	50
Crime and Punishment	4	46
The Gambler	5	11

3.7.4 The user gives their opinion

A common usability principle is that it is easier for humans to recognize items, than to draw them from memory. Therefore, it is sometimes easier for a user to say what they want or do not want, when they have options in front of them. The options mentioned can be simplified to be mutually exclusive, e.g. either a user likes an item or they do not. It is equally possible to create an explanation facility using a sliding scale.

In previous work in recommender systems a user could for example specify whether they think an item is interesting or not, if they would like to see more similar items, or if they have already seen the item previously (Billsus and Pazzani, 1999; Swearingen and Sinha, 2002).

3.7.5 Mixed interaction interfaces

Chen and Pu (2007) evaluated an interface where users could both specify their requirements from scratch, or make alterations to existing requirements generated by the system (a dynamic critiquing interface, as described in Section 2.1.3), and found that the mixed interface increases efficiency, satisfaction as well as effectiveness of decisions compared to only making alterations to existing sets of requirements.

McNee et al. (2003b) evaluated a hybrid interface for rating items. In this study users were able to rate items suggested by the system, as well as search for items to rate themselves. The mixed-initiative model did not outperform either rating model in terms of the accuracy of resulting recommendations. On the contrary, allowing users to search for items to rate increased loyalty (measured in terms of returns to the system, and number of items rated).

3.8 Explanation styles (per algorithm)

Table 3.4: Examples of explanations in commercial and academic systems, ordered by explanation style (case-based, collaborative, content, conversational, demographic and knowledge/utility-based.)

System	**Example explanation**	**Explanation style**
iSuggest-Usability (Hingston, 2006)	See e.g. Figure 3.11	Case-based
LoveFilm.com	*"Because you have selected or highly rated: Movie A"*	Case-based
LibraryThing.com	*"Recommended By User X for Book A"*	Case-based
Netflix.com	A list of similar movies the user has rated highly in the past	Case-based
Amazon.com	*"Customers Who Bought This Item Also Bought . . . "*	Collaborative
LIBRA (Bilgic and Mooney, 2005)	Keyword style (Tables 4.3 and 4.4); Neighbor style (Figure 4.3); Influence style (Figure 3.3)	Collaborative
MovieLens (Herlocker et al., 2000)	Histogram of neighbors (Figure 3.2) and Confidence display (Figure 3.6)	Collaborative
Amazon.com	*"Recommended because you said you owned Book A"*	Content-based
CHIP (Cramer et al., 2008b)	*"Why is 'The Tailor's Workshop recommended to you'? Because it has the following themes in common with artworks that you like: * Everyday Life * Clothes . . . "*	Content-based
Moviexplain (Symeonidis et al., 2008)	See Table 3.5	Content-based
MovieLens: *"Tagsplanations"* (Vig et al., 2009)	Tags ordered by relevance or preference (see Figure 3.10)	Content-based
News Dude (Billsus and Pazzani, 1999)	*"This story received a [high/low] relevance score, because it contains the words f1, f2, and f3."*	Content-based

OkCupid.com	Graphs comparing two users according to dimensions such as “more introverted”; comparison of how users have answered different questions	Content-based
Pandora.com	*“Based on what you’ve told us so far, we’re playing this track because it features a leisurely tempo . . . ”*	Content-based
Adaptive place Advisor (Thompson et al., 2004)	Dialog e.g. “Where would you like to eat?” “Oh, maybe a cheap Indian place.”	Conversational
ACORN (Wärnestål, 2005b)	Dialog e.g. *“What kind of movie do you feel like?” “I feel like watching a thriller.”*	Conversational
INTRIGUE (Adrissono et al., 2003)	*“For children it is much eye-catching, it requires low background knowledge, it requires a few seriousness and the visit is quite short. For yourself it is much eye-catching and it has high historical value. For impaired it is much eye-catching and it has high historical value.”*	Demographic
Qwikshop (McCarthy et al., 2004)	*“Less Memory and Lower Resolution and Cheaper”*	Knowledge/utility-based
SASY (Czarkowski, 2006)	*“. . . because your profile has: *You are single; *You have a high budget”* (Figure 3.1)	Knowledge/utility-based
Top Case (McSherry, 2005)	*“Case 574 differs from your query only in price and is the best case no matter what transport, duration, or accommodation you prefer”*	Knowledge/utility-based
(Internet Provider) (Felfernig and Gula, 2006)	*“This solution has been selected for the following reasons: *Webspace is available for this type of connection . . . ”* (Figure 3.3)	Knowledge/utility-based

"Organizational Structure" (Pu and Chen, 2006)	Structured overview: *"We also recommend the following products because: *they are cheaper and lighter, but have lower processor speed."* (Figure 3.8)	Knowledge/utility-based
myCameraAdvisor (Wang and Benbasat, 2007)	e.g *"... cameras capable of taking pictures from very far away will be more expensive ... "*	Knowledge/utility-based

Table 3.4 summarizes the most commonly used explanation styles (case-based, content-based, collaborative-based, demographic-based, knowledge and utility-based with examples of each. In this section we describe each style: their corresponding inputs, processes and generated explanations. For commercial systems where this information is not public, we offer educated guesses. While conversational systems are included in the Table, we refer to Section 3.7.1 which refers to explanations in conversational systems as more of an interaction style than a particular algorithm.

The explanation style for a given explanation may, or may not, reflect the underlying algorithm by which they are computed. That is to say that the explanations may also follow the "style" of a particular algorithm irrespective of whether or not this is how they have been retrieved/computed. For example, it is possible to have a content-based explanation for a recommendation engine using collaborative filtering. Consequently this type of explanation would not be consistent with the criterion of transparency, but may support other explanatory criteria. In the following sections we will give further examples of how explanation styles can be inspired by these common algorithms.

For describing the interface between the recommender system and explanation component we use the notation used in Burke (2002): **U** is the set of users whose preferences are known, and **u** $\in U$ is the user for whom recommendations need to be generated. **I** is the set of items that can be recommended, and **i** $\in$ **I** is an item for which we would like to predict u's preferences.

3.8.1 Collaborative-Based Style Explanations

For collaborative-based style explanations the assumed input to the recommender engine are user **u**'s ratings of items in **I**. These ratings are used to identify users that are similar in ratings to **u**. These similar users are often called "neighbors" as nearest-neighbors approaches are commonly used to compute similarity. Then, a prediction for the recommended item is extrapolated from the neighbors' ratings of **i**.

Commercially, the most well known usage of collaborative-style explanations are the ones used by Amazon.com: *"Customers Who Bought This Item Also Bought ... "*. This explanation assumes that the user is viewing an item which they are already interested in. The system finds similar users (who bought this item), and retrieves and recommends items that similar users bought. The recommendations are presented in the format of similar to top item. In addition, this explanation assumes an interaction model, whereby ratings are implicitly inferred through purchase behavior.

Herlocker et al. suggested 21 explanation interfaces using text as well as graphics (Herlocker et al., 2000). These interfaces varied with regard to content and style, but a number of these explanations directly referred to the concept of neighbors. Figure 3.2 for example, shows how neighbors rated a given (recommended) movie, a bar chart with "good", "ok" and "bad" ratings clustered into distinct columns. Again, we see that this explanation is given for a specific way of recommending items, and a particular interaction model: this is a single recommendation (either top item or one item out of a top-N list), and assumes that the users are supplying rating information for items.

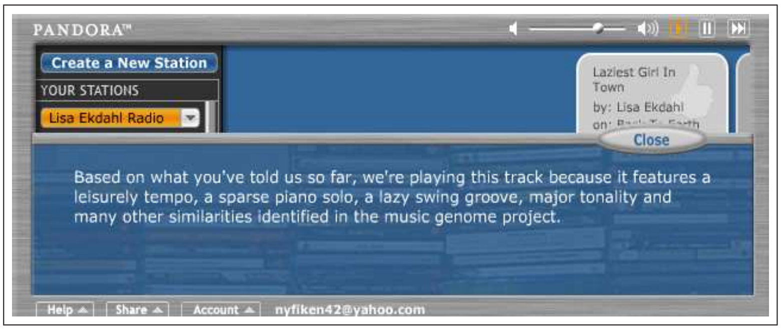
3.8.2 Content-Based Style Explanation

For content-based style explanations the assumed input to the recommender engine are user **u**'s ratings (for a sub-set) of items in **I**. These ratings are then used to generate a classifier that fits **u**'s rating behavior and use it on **i**. A prediction for the recommended item is based on how well it fits into this classifier. E.g. if it is similar to other highly rated items.

If we simplify this further, we could say that content-based algorithms consider similarity between items, based on user ratings but considering item properties. In the same spirit, content-based style explanations are based on the items' properties. For example, Symeonidis et al. (2008) justify a movie recommendation according to what they infer is the user's favorite actor (see Table 3.5). While the underlying approach is in fact a hybrid of collaborative and content-based approaches, the explanation style suggests that they compute the similarity between movies according to the presence of features in highly rated movies. They elected to present users with several recommendations and explanations (top-N) which may be more suitable if the user would like to make a selection between movies depending on the information given in the explanations (e.g. feeling more like watching a movie with Harrison Ford over one starring Bruce Willis). The interaction model is based ratings of items.

A more domain independent approach is suggested by Vig et al. (2009) who use the relationship between tags and items (tag relevance) and the relationship between tags and users (tag preference) to make recommendations (see Figure 3.10). Tag preference can be seen as a form of content-based explanation, as it is based on a user's ratings of movies with that tag. Here, showing recommendations as a single top item allows the user to view many of the tags that are related to the item. The interaction model is again based on numerical ratings.

The commercial system Pandora, explains its recommendations of songs according to musical properties such as tempo and tonality. These features are inferred from users ratings of songs. Figure 3.9 shows an example of this [6]. Here, the user is offered one song at a time (top item) and gives their opinion as "thumbs-up" or "thumbs-down" which also can be considered as numerical ratings.



Based on what you've told us so far, we're playing this track because it features a leisurely tempo, a sparse piano solo, a lazy swing groove, major tonality and many other similarities identified in the music genome project.

Figure 3.9: Pandora explanation: *"Based on what you've told us so far, we're playing this track because it features a leisurely tempo . . . "*

Table 3.5: Example of an explanation in Moviexplain, using features such as actors, which occur for movies previously rated highly by this user, to justify a recommendation (Symeonidis et al., 2008)

Recommended movie title	**The reason is the participant**	**who appears in**
Indiana Jones and the Last Crusade (1989)	Ford, Harrison	5 movies you have rated
Die Hard 2 (1990)	Willis, Bruce	2 movies you have rated

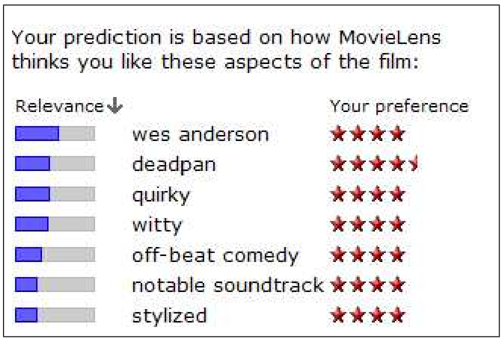
Case-Based Style Explanations

A content-based explanation can also omit mention of significant properties and focus primarily on the items used to make the recommendation. The items used are thus considered cases for comparison, resulting in case-based style explanations.

In this chapter we have already seen a type of case-based style explanation, the "influence based style explanation" of Bilgic and Mooney (2005) in Table 3.3. Here, the influence of an item on the recommendation is computed by looking at the difference in the score of the recommendation with and without that item. In this case, recommendations were presented as top item, assuming a rating based interaction. Hingston (2006) computed the similarity between recommended items[7], and used these similar items as

[6] http://www.pandora.com - retrieved Nov. 2006

[7] The author does not specify which similarity metric was used, though it is likely to be a form of rating based similarity measure such as cosine similarity.



Your prediction is based on how MovieLens thinks you like these aspects of the film:

Relevance ↓		Your preference
▬	wes anderson	★★★★
▬	deadpan	★★★★½
▬	quirky	★★★★
▬	witty	★★★★
▬	off-beat comedy	★★★★
▬	notable soundtrack	★★★★
▬	stylized	★★★★

Figure 3.10: Tagsplanation with both tag preference and relevance, but sorted by tag relevance

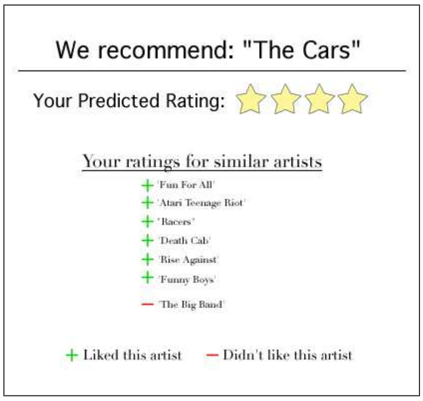
justification for a top item recommendation in the "learn by example" explanations (see Figure 3.11).

3.8.3 Knowledge and Utility-Based Style Explanations

For knowledge and utility-based style explanations the assumed input to the recommender engine are description of user **u**'s needs or interests. The recommender engine then infers a match between the item **i** and **u**'s needs. One knowledge-based recommender system takes into consideration how camera properties such as memory, resolution and price reflect the available options as well as a user's preferences (McCarthy et al., 2004). Their system may explain a camera recommendation in the following manner: *"Less Memory and Lower Resolution and Cheaper"*. Here recommendations are presented as a form of structured overview describing the competing options, and the interaction model assumes that users ask for alterations in the recommended items.

Similarly, in the system described in McSherry (2005) users gradually specify (and modify) their preferences until a top recommendation is reached. This system can generate explanations such as the following for a recommended holiday titled "Case 574": *"Top Case: Case 574 differs from your query only in price and is the best case no matter what transport, duration, or accommodation you prefer"*.

It is arguable that there is a certain degree of overlap between knowledge-based and content-based explanations. This is particularly the case for case-based style explanations (Section 3.8.2) which can be derived from either type of algorithm depending on the details of the implementation.



We recommend: "The Cars"

Your Predicted Rating: ★★★★

Your ratings for similar artists

+ 'Fun For All'
+ 'Atari Teenage Riot'
+ 'Racers'
+ 'Death Cab'
+ 'Rise Against'
+ 'Funny Boys'
− 'The Big Band'

+ Liked this artist − Didn't like this artist

Figure 3.11: Learn by example, or case based reasoning, Hingston (2006)

3.8.4 Demographic Style Explanations

For demographic-based style explanations, the assumed input to the recommender engine is demographic information about user **u**. From this, the recommendation algorithm identifies users that are demographically similar to **u**. A prediction for the recommended item **i** is extrapolated from how the similar users rated this item, and how similar they are to **u**.

Surveying a number of systems which use a demographic-based filter (e.g. Adrissono et al., 2003; Krulwich, 1997; Pazzani, 1999), we could only find one which offers an explanation facility: *"For children it is much eye-catching, it requires low background knowledge, it requires a few seriousness and the visit is quite short. For yourself it is much eye-catching and it has high historical value. For impaired it is much eye-catching and it has high historical value."*(Adrissono et al., 2003). In this system recommendations were offered as a structured overview, categorizing places to visit according to their suitability to different types of travelers (e.g. children, impaired). Users can then add these items to their itinerary, but there is no interaction model that modifies subsequent recommendations

To our knowledge, there are no other systems that make use of demographic style explanations. It is possible that this is due to the sensitivity of demographic information; anecdotally we can imagine that many users would not want to be recommended an item based on their gender, age or ethnicity (e.g. *"We recommend you the movie Sex in the City because you are a female aged 20-40."*).

3.9 Summary

In this chapter, we offer guidelines for the designers of explanations in recommender systems. Firstly, the designer should consider what benefit the explanations offer, and thus which criteria they are evaluating the explanations for (e.g. *transparency, scrutability,* *trust, efficiency, effectiveness, persuasion or satisfaction*). The developer may select several criteria which may be related to each other, but may also be conflicting. In the latter case, it is particularly important to distinguish between these evaluation criteria. It is only in more recent work that these trade-offs are being shown and becoming more apparent (Cramer et al., 2008b; Tintarev and Masthoff, 2008b).

In addition, the system designer should consider the *metrics* they are going to use when evaluating the explanations, and the dependencies the explanations may have with different parts of the system, such as the way recommendations are presented (e.g. top item, top N-items, similar to top item(s), predicted ratings for all items, structured overview), the way users interact with the explanations (e.g. the user specifies their requirements, asks for an alteration, rates items, gives their opinion, or uses a hybrid interaction interface) and the underlying recommender engine.

To offer a single example of the relation between explanations and other recommender system factors, we can imagine a recommender engine with low recommendation accuracy. This may affect all measurements of effectiveness in the system, as users do not really like the items they end up being recommend. These measurements do not however reflect the effectiveness of the *explanations* themselves. In this case, a layered approach to evaluation (Paramythis et al., 2001), where explanations are considered in isolation from the recommendation algorithm as seen in Tintarev and Masthoff (2008b) (see also Chapter 6), may be warranted. Similarly, thought should be given how the method of presenting recommendations, and the method of interaction may affect the (evaluation of) explanations.

We offered examples of the most common explanation styles from existing systems, and explain how these can be related to the underlying algorithm (e.g. content-based, collaborative, demographic, or knowledge/utility-based). To a certain extent these types of explanations can be reused (likely at the cost of transparency) for hybrid recommendations, and other complex recommendation methods such as latent semantic analysis, but these areas of research remain largely open. Preliminary works for some of these areas can be found in (e.g. Khan et al., 2009; Hu et al., 2008) (explaining Markov decision processes and latent semantic analysis models).

In this thesis, we will particularly consider the aim of effectiveness as the work on this particular criterion has been limited. In any evaluation of a given criterion it is however important to realize that this is one of many possible criteria, and it is worthwhile to study the trade-offs involved with optimizing on a single criterion. Later in this thesis we will consider how using explanation content that is personalized to participants affects primarily the criterion of effectiveness, but also persuasion and satisfaction (Chapters 6 and 7). The next chapter (Chapter 4) is dedicated to a further elaboration of the definition and measurement of the criterion of effectiveness, and in particular the effect personalization may have on this metric.

Chapter 4

Effectiveness and personalization

4.1 Introduction

In the previous chapter we discussed different explanatory aims, and how they relate to factors such as the degree of interaction and how recommendations are presented. In this chapter we narrow down the discussion to the particular explanatory aim of effectiveness, or to explanations that assist users in making *correct* decisions. Even more specifically, we elect to discuss the role of personalization on effectiveness. In Section 4.2, we review previous evaluations which form an argument for studying the effect of personalization on effectiveness.

Next, we survey the ways in which effectiveness can be measured in Section 4.3. In Section 4.3.1 we define effectiveness as a metric which we adapt and reuse in our empirical studies in Chapters 6 and 7. This metric is a difference between two item ratings (*Rating1-Rating2*) where:

1. **(Rating1)** The user rates the product on the basis of the explanation
2. The user tries the product (alternatively approximate by reading online reviews)
3. **(Rating2)** The user re-rates the product

Since the metric of effectiveness may be domain dependent, in Section 4.5 we discuss an exploratory experiment in which we measure *perceived* effectiveness in different product domains. While effectiveness measures whether the decisions made by users were truly correct, *perceived* effectiveness is the users' subjective evaluation of whether the explanations offer them the information they need to make good decisions. *Perceived* effectiveness is different from persuasion in that no influence on the user is intended and a true evaluation is assumed, but it is naturally still an approximation of true effectiveness and limited as such.

As we mentioned in previous chapters, the underlying recommendation algorithm can shape the explanation content. Therefore, in Section 4.6 we briefly summarize the results of other exploratory studies on presentational choices and *explanation styles* including those influenced by different underlying algorithms: collaborative-based, content-based and case-based filtering. Our thoughts on effectiveness are summarized in Section 4.7.

4.2 An argument for considering personalization for effective explanations

4.2.1 Overview

As mentioned in the previous chapter, the focus of this thesis is on explanations which aim to help users make qualified decisions, i.e. effective explanations. We start this chapter with an overview of an argument for considering personalization for effective explanations in recommender system, based on previous user studies. We highlight the previous studies which we consider most relevant, and dedicate a subsection to each. Although we will return to these studies shortly, we first summarize their combined contribution.

The seminal experiment by Herlocker et al. (2000) compared different explanation interfaces for a movie recommender, measuring how likely a user thought they would be to see this movie at the cinema. Although the authors argue that they measure effectiveness, this is more a measure of persuasion than effectiveness. Their results also suggest that item specific features such as favourite actor/actress might vary in the importance they carry from user to user. The explanations used in Herlocker et al. (2000)'s are varied in terms of both content and presentation, but the experiment has several limitations which we address in our own work. A more detailed review of this study can be found in Section 4.2.2

Similarly to Herlocker et al. (2000), Bilgic and Mooney (2005) compared different types of explanations, but this time for effectiveness. The explanations were evaluated according to how much they helped users find books they still liked after receiving more detailed information (effectiveness). This work brings up two interesting points: a) there is a difference between persuasion - the user thinking they would like the item and effectiveness - helping the user decide whether they truly like the item and b) that the winning explanation found by Herlocker et al. (2000) is likely to cause overestimation of items.

Hingston (2006) studied the perceived effectiveness of explanations for a number of interfaces and recommendation algorithms. For the case-based explanation interfaces (which compared the recommended item to similar items the user liked in the past), participants requested information about why items were judged to be similar to one another.

This is a similar result to Bilgic and Mooney (2005) who failed to show a significant result on effectiveness for an explanation interface which used information about previously rated items, but where the explicit relations between these previously rated items and the current recommendation were not clear.

The results from the studies of Bilgic and Mooney (2005) and Hingston (2006) suggest that it is not enough for participants to be told that two items are similar, and that an explanation facility may benefit from describing similarity in terms of shared item features. However, considering particular features may be more important for explanations aiming at effectiveness rather than other criteria. It can for example be imagined that information about recommendation confidence or previous performance (two interfaces used in the study by Herlocker et al. (2000)), could be used to gain user trust in the recommender engine.

Other related studies are those of Carenini and Moore (2000a, 2001), who looked at evaluative arguments for house recommendations, and like Herlocker et al. (2000) measured persuasion. However, the generated arguments also presented negative information about the recommended houses, and may have rendered interesting results for an evaluation of effectiveness. They found personalized evaluative arguments to be more persuasive in the house domain (Carenini and Moore, 2001). This body of work is discussed in Section 4.2.5. Their results inspired the question of whether personalization of explanations may help optimize for the aim of effectiveness as well. Here we also note that the choice of houses as a domain, while interesting from the perspective of user involvement due to a high financial investment, was limited by the fact that they could not measure the actual acceptance of items.

We build upon the preceding research, by considering that using personalized item features in explanations may help increase effectiveness in Chapters 6 and 7.

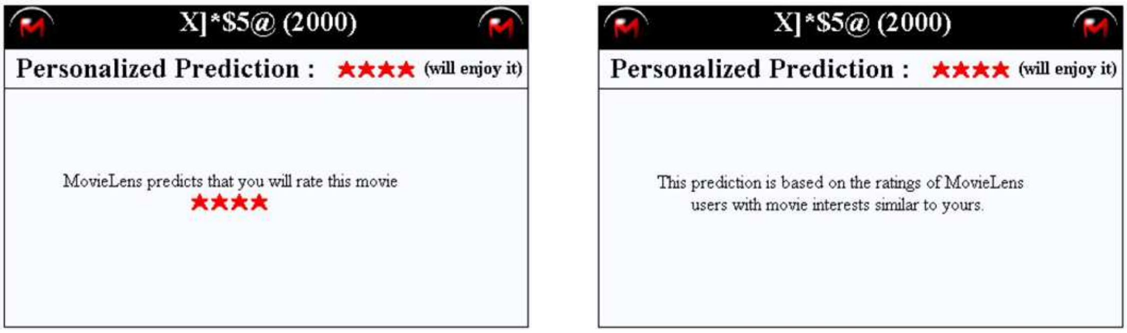
The content of this thesis is novel in that it offers a more thorough study for the criterion of effectiveness, considering several factors such as presentation and personalization. In the following sections we will elaborate on the related studies we have just mentioned. If the reader is satisfied with the brief overview given above, we recommend passing over these and resuming the reading of this chapter in Section 4.3 where we discuss metrics for evaluating effectiveness.

4.2.2 A review of Herlocker et al. (2000) - evaluating explanation interfaces

Herlocker et al. evaluated twenty-one different explanation interfaces in the context of a movie recommender system - MovieLens (Herlocker et al., 2000), measuring how likely a user thought they would be to see this movie at the cinema. Although the authors argue that they measure effectiveness, this is more a measure of persuasion than effectiveness.

We return to this important distinction between persuasion and effectiveness, as highlighted by Bilgic and Mooney (2005), in Section 4.2.3. Their results also suggest that item specific features such as favourite actor/actress might vary in the importance they carry from user to user.

The underlying algorithm of MovieLens is collaborative filtering and is based on similarity between users' ratings of movies on a 1-5 star scale. Each user was asked to evaluate their likelihood of seeing a particular movie for each of the twenty-one interfaces. The same movie was recommended in each explanation, and was based on a recommendation for the primary author of the paper. The explanations were therefore the same for each participant, and not personalized. However, the title of this movie was encoded and was thus unknown to participants. Out of the seven aims of explanations in recommender systems suggested by this thesis, the mentioned experiment would be aiming to persuade users to watch the movie, as we have no information about the user's genuine evaluation of the movie (effectiveness). Table 4.1, summarizes their results for each type of explanation on a scale from one to seven. Explanations 11 (Figure 4.1a) and 12 (Figure 4.1b) represent the two base cases of explanations with no additional information.



X]*$5@ (2000)

Personalized Prediction : ★★★★ (will enjoy it)

MovieLens predicts that you will rate this movie ★★★★

(a) Focus on system - "Movie lens predicts that you will rate this movie four stars"

X]*$5@ (2000)

Personalized Prediction : ★★★★ (will enjoy it)

This prediction is based on the ratings of MovieLens users with movie interests similar to yours.

(b) Focus on users - "This prediction is based on the ratings of MovieLens users with movie interests similar to yours".

Figure 4.1: The two explanation with no additional information, used as baselines by Herlocker (2000)

The explanation interfaces varied both in terms of presentation and content. Some of the interfaces were graphical, others textual. The content of explanation varied from explanations based on the collaborative-based engine (e.g. "similar users", see also Figure 4.2 a), others describe features of the the movies, e.g. favourite actor/actress, yet a third category of explanations say other things about the recommendation such as how confident the engine is about its recommendation etc.

Participants were most likely to see the movie if they saw a histogram of how similar users had rated the item, with one bar for "good" and another for "bad" ratings (see Figure 4.2a). However, we argue that a weakness of this result is a bias toward positive ratings

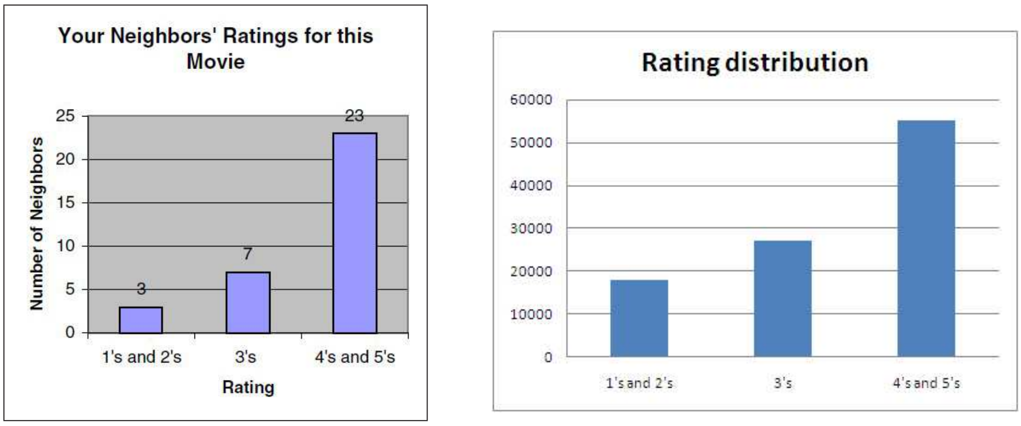
Table 4.1: Mean response of users to each explanation interface, based on a scale of one to seven (where seven means most likely to see the movie). Explanations 11 and 12 (see Figures 4.1a and b) represent the base cases of no additional information. Shaded rows indicate explanations with a mean response significantly different from the base cases (two-tailed $p <= 0.05$). Screen shots of all interfaces can be found in (Herlocker, 2000).

Ranking	Description	Mean (StD)
1	Similar users (Histogram with grouping)	5.25 (1.29)
2	Past Performance	5.19 (1.16)
3	Neighbor ratings histogram	5.09 (1.22)
4	Table of neighbors ratings	4.97 (1.29)
5	Similarity to other movies rated	4.97 (1.50)
6	Favorite actor/actress	4.92 (1.73)
7	Confidence in prediction	4.71 (1.02)
8	Won awards	4.67 (1.49)
9	Detailed process description	4.64 (1.40)
10	# neighbors	4.60 (1.29)
11	No extra data - focus on system	4.53 (1.20)
12	No extra data - focus on user	4.51 (1.35)
13	MovieLens confidence in prediction	4.51 (1.35)
14	Good profile	4.45 (1.53)
15	Overall percent rated 4+	4.37 (1.26)
16	Complex graph: count, ratings, similarity	4.36 (1.47)
17	Recommended by movie critics	4.21 (1.47)
18	Rating and % agreement of closest neighbor	4.21 (1.20)
19	# neighbors with std. deviation	4.19 (1.45)
20	# neighbors with avg. correlation	4.08 (1.46)
21	Overall average rating	3.94 (1.22)

in the MovieLens dataset. Figure 4.2b) illustrates the distribution of ratings in the public MovieLens dataset of 100.000 ratings. We see that a very large proportion of ratings in the data-base are high (4's or 5's). 80% of ratings are 3 or higher, and 55% are 4's or 5's. Although computing similarity to other user ratings is likely to give a more accurate estimate than metrics such as simple average, such a severe skew in the underlying data is likely to result in over-estimates of ratings overall. Bilgic and Mooney (2005) show that this type of explanation may lead to overestimation (see also Section 4.2.3).

There was relatively poor acceptance for explanations using information about the user's favourite actor or actress. We argue however that this is another weakness of the experiment - it would seem plausible that this property (favourite actor/actress) is more important to some users than others. This intuition is backed up by qualitative feedback we received in focus groups and analysis of online movie reviews (see Chapter 5), as well as the high variance in acceptance for this type of explanations in Herlocker et al. (2000)'s experiment. It is also noteworthy that in the experiment no actor names were mentioned, only that they were the user's favorite.

This experiment is an exhaustive piece of work in the area of explanations in recommender systems - in particular in terms defining which content to present. We wish to extend this work in a number of ways. Firstly, we would like to see how these explanation interfaces fare in an evaluation measuring effectiveness rather than persuasion. Secondly, we would be interested to study the effects of using explanation content that is personalized to participants. Finally, while the experiment is limited to explanations that can be generated in MovieLens, and thus using a collaborative-based algorithm, our study will not be restricted by an underlying algorithm.



(a) Leading histogram

(b) Distribution of ratings in MovieLens

Figure 4.2: Skew of ratings in MovieLens 100K rating dataset compared to Herlocker et al. (2000)'s leading histogram

4.2.3 A review of (Bilgic and Mooney, 2005) - the difference between persuasion and effectiveness

Bilgic and Mooney (2005) argue that there is a difference between the two aims of explanation which in our work we call persuasion and effectiveness respectively. That is, that while some explanations may manage to convince users to try or buy a recommended item (persuasion), they may fail at actually helping users make more accurate decisions (effectiveness). For this purpose, Bilgic and Mooney (2005) compare how three different types of explanation interfaces help users assess the actual rating for items. They also argue that the winning explanation found by Herlocker et al. (2000) is likely to cause overestimation of items.

In their experiment Bilgic and Mooney (2005) evaluated the effectiveness of three explanation styles. The working assumption was that the definition of an effective explanation is to allow a user to correctly assess the item. Due to time restrictions participants did not have the possibility to fully try the items (read the books). Rather, trying the item was approximated by allowing participants to base their second rating on information available on Amazon's website.

The three interfaces used are depicted in Table 4.2, Figure 4.3, and Tables 4.3 and 4.4. Table 4.2 depicts an influence based explanation which shows which ratings influenced the recommended item (a book) the most. Note that the neighbour style explanation in Figure 4.3 is very similar to the wining histogram in the study by Herlocker et al. (2000) (Figure 4.2). The third interface is a keyword style explanation (Table 4.3), where the user can see which keywords contributed to their recommendation. For this type of explanation users can also ask for a more detailed explanation of how the strength of each keyword was computed (in terms of items they had previously rated) by clicking on *"Explain"* (Table 4.4 shows an example of this expansion).

Bilgic and Mooney (2005) found the keyword style explanation (Tables 4.3 and 4.4) to be effective, but the influence based (Table 4.2) or neighborhood style (Figure 4.3) explanations were not effective. In fact, the neighborhood style explanation, also used in Herlocker et al. (2000)'s study, described in 4.2.2) caused users to overestimate the actual item rating. The skew towards positive ratings in the recommendations (see also our argument in Section 4.2.2) is likely to have caused this overestimation. We hypothesize that

Table 4.2: Influence of previous book ratings, on the current book recommendation (Bilgic and Mooney, 2005)

BOOK	**YOUR RATING Out of 5**	**INFLUENCE Out of 100**
Of Mice and Men	4	54
1984	4	50
Till We Have Faces: A Myth Retold	5	50
Crime and Punishment	4	46
The Gambler	5	11

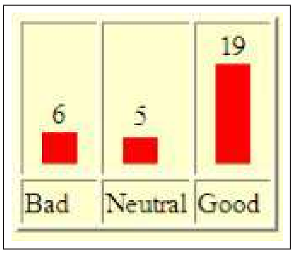


Figure 4.3: The Neighbor Style Explanation - a histogram summarizing the ratings of similar users (neighbors) for the recommended item grouped by good (5 and 4's), neutral (3s), and bad (2s and 1s), on a scale from 1 to 5. The similarity to Figure 3.2 in this study was intentional, and was used to highlight the difference between persuasive and effective explanations Bilgic and Mooney (2005).

Table 4.3: The keyword style explanation by Bilgic and Mooney (2005). This recommendation is explained in terms of keywords that were used in the description of the item, and that have previously been associated with highly rated items. “Count” identifies the number of times the keyword occurs in the item’s description, and “strength” identifies how influential this keyword is for predicting liking of an item.

Slot	**Word**	**Count**	**Strength**	**Explain**
DESCRIPTION	HEART	2	96.14	*Explain*
DESCRIPTION	BEAUTIFUL	1	17.07	*Explain*
DESCRIPTION	MOTHER	3	11.55	*Explain*
DESCRIPTION	READ	14	10.63	*Explain*
DESCRIPTION	STORY	16	9.12	*Explain*

Table 4.4: A more detailed explanation for the “strength” of a keyword which shows after clicking on *“Explain”* in Table 4.3. In practice “strength” probabilistically measures how much more likely a keyword is to appear in a positively rated item than a negatively rated one. It is based on the user’s previous positive ratings of items (“rating”), and the number of times the keyword occurs in the description of these items (“count”) Bilgic and Mooney (2005).

Title	**Author**	**Rating**	**Count**
Hunchback of Notre Dame	Victor Hugo, Walter J. Cobb	10	11
Till We Have Faces: A Myth Retold	C.S. Lewis, Fritz Eichenberg	10	10
The Picture of Dorian Gray	Oscar Wilde, Isobel Murray	8	5

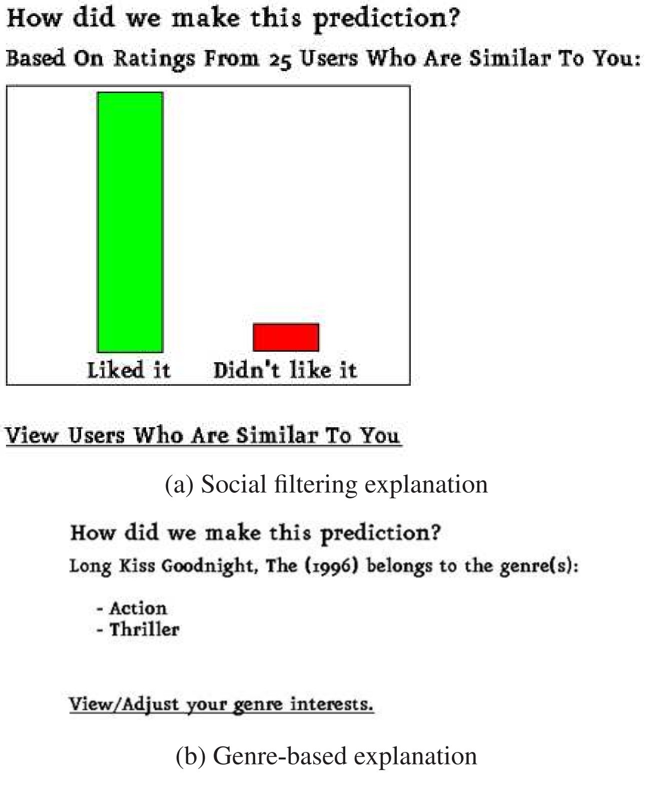
the good result for keyword style explanations is due to the keywords helping users better understand what it is about the previously rated items that is similar to the recommended item - while no such explicit information exists for the influence style (or neighbourhood style) explanations.

This study has therefore raised two important issues, firstly the importance of differentiating between persuasion and effectiveness, and secondly, the overestimation that a neighborhood style explanation may cause. The results of this work also suggest that it may be worthwhile to investigate if explanations based on similarity between particular item keywords or features could be most effective. That is, even if the underlying algorithm selects items according to another similarity measure (e.g. similar users), the most effective explanation may still be one based on keywords or features.

4.2.4 A review of Hingston (2006) - a case for using item features in explanations

Among other things, in his honour's thesis, Hingston (2006) studied the *perceived* usefulness (effectiveness) and understandability of different explanation interfaces, based on a variety of underlying algorithms. Each recommendation was accompanied by a short explanation of how the recommendation method selected this item, and participants were asked questions such as how useful and understandable they perceived the explanations to be, and to rank explanations in order of usefulness. These assessments by users give some idea of the effectiveness of the explanations, but they are not equivalent to measuring true effectiveness such as described in 4.3.

Hingston (2006)'s studies suggest a correlation between an explanation's understandability and its degree of perceived usefulness. For the most part, participants felt that explanations added to the usefulness of the recommendations. It was also clear that explanation facilities have the potential to both decrease and increase the perceived usefulness of a recommendation depending on how easy they are to understand. In this study, the two interfaces that were considered the most useful were based on social filtering and (movie) genre (Figures 4.4a and b respectively). The social explanations are similar to those in the work of Herlocker et al. (2000), described in Section 4.2.2, and suffer from the same limitations; e.g. a bias toward positive ratings.


How did we make this prediction?

Based On Ratings From 25 Users Who Are Similar To You:

View Users Who Are Similar To You

(a) Social filtering explanation

How did we make this prediction?

Long Kiss Goodnight, The (1996) belongs to the genre(s):

- Action
- Thriller

View/Adjust your genre interests.

(b) Genre-based explanation

Figure 4.4: Two explanation interfaces taken from the study by Hingston (2006)

Noteworthy qualitative comments from participants include requests for additional information in explanations. For the genre based explanations such as in Figure 4.4b, the participants wanted to know how much the system assumes they were interested in each genre e.g. a participant might want to know how strongly the system represents their interests for the respective genres "Action" and "Thriller". For the case-based explanations which explained by comparing to similar items the user liked in the past (Figure 3.11), participants requested information about why items were judged to be similar to one another. For this explanation type participants also suggested that the user should be able to adjust the factors that are used to judge similarity between items. This study therefore raises the question if whether considering particular item features, and weighing the factors (e.g. actor or genre) for computing similarity between items, when constructing explanations, are not only things that users want, are also things that increases the effectiveness of explanations.

4.2.5 A review of Carenini (2001) - a case for personalization

Carenini (2001) conducted his PhD thesis on the topic of evaluative arguments, that is, communicative acts that attempt to advise or persuade the hearer that something is good (vs. bad) or right (vs. wrong). They claim to measure the effectiveness of personalized evaluative arguments, that presented both negative and positive information, and found that personalized arguments were more effective than non-personalized (Carenini and Moore, 2001). For effectiveness they measured the acceptance of recommended items, which cannot be said to be (true) effectiveness however, because the choice domain (houses) was limited by the fact that they could not measure the true liking of items.

The underlying system (GEA - Generator of Evaluative Arguments) is based on a natural language generation framework, fortified by argumentation theory, and generates user tailored arguments for buying houses. An example of the arguments that were generated by GEA is: *"House 3-26 is an interesting house, in fact, it has a convenient location in the safe Eastend neighborhood. Even though house 3-26 is somewhat far from a rapid transportation stop (1.6 miles), it is close to work (1.8 miles)..."*

The tailoring is based on multi-attribute utility theory (MAUT) - weighing the arguments according to their importance to the user, as well as which features are most "compelling" - defined according to how much they distinguish between items. Although this work is focused on evaluative text rather than explanations, we believe that evaluative arguments are likely to help users make informed decisions, i.e. improve explanation effectiveness. In addition, being a natural language generation system, this work addresses relevant questions in regarding to optimizing natural language, such as the optimal degree of conciseness, and the role of tailoring text to the user. In several task-based evaluations, Carenini (2001) asked participants to make a selection of a subset of preferred items (houses), and then asked for self reports of participant appreciation of the items in their selection. Since these reports cannot reveal user's true appreciation of the houses, but only their appreciation of their recommendations, these evaluations measured the effect of the arguments on persuasion. However, as the generated arguments are not limited to positive information about the items, an evaluation of effectiveness could also be imagined.

In these experiments, they found that tailored arguments were more persuasive than non-tailored (Carenini and Moore, 2001). They also found that concise arguments were more persuasive than verbose (Carenini and Moore, 2000a). In this thesis, we will therefore also consider the effect of tailoring explanations with regard to the aim of effectiveness (Chapter 6), and very briefly consider the effect of modifying explanation length in a pilot study described in Section 5.4.

4.3 Evaluating effectiveness

In Chapter 3, we briefly touched upon metrics for different explanatory aims. As we have previously stated, this thesis is focused on evaluating explanations for effectiveness, or decision support. In this section, we delve deeper into how effectiveness can be measured.

4.3.1 Gives correct valuation of item

One way to measure decision support for explanations in recommender systems is suggested in Bilgic and Mooney (2005), using the following steps:

1. **(Rating1)** The user rates the product on the basis of the explanation
2. The user tries the product (alternatively approximate by reading online reviews)
3. **(Rating2)** The user re-rates the product

Effectiveness can then be measured by the discrepancy between Steps 1 and 3 ($Rating1 - Rating2$). Bilgic and Mooney (2005) approximated Step 2, by letting the users view reviews of the items (books) online. For other domains, other types of approximation may be relevant, such as trailers for movies. According to this metric, an effective explanation is one which minimizes the gap between these two ratings. If an explanation helps users make good decisions, getting more (accurate and balanced) information or trying the product should not change their valuation of the product greatly.

The difference between the two ratings may be positive (overestimation of the product) or negative (underestimation). Overestimation will lead to more false positives; users trying products they end up liking less than they anticipated. Particularly in high investment recommendation domains such as holidays, a false positive may result in a large blow to trust in the system. Underestimation will on the other hand lead to more false negatives; user missing products they might have appreciated. If a user recognizes an underestimation due to previous knowledge or subsequent exposure, this may lead to a loss of trust as well. An underestimation may also needlessly decrease an e-commerce site's revenue.

Bilgic and Mooney (2005) looked at the mean of this difference of these two ratings, with 0 being the best possible mean. In a normal distribution, with as much overestimation as underestimation, the mean effectiveness will be close to 0, but this does not mean the explanations are effective. Bilgic and Mooney (2005) remedy this by a complimentary measure - *also* looking at the correlation between the first and second rating, with a high and significant correlation implying high significance.

4.3.2 Helps user find the best item

In some cases, such as in high investment domains, it may be more relevant for a user to find the best possible item. Chen and Pu (2007) evaluated explanation interfaces in a number ways including effectiveness. In their study, participants interacted with a recommender system, but with the aim to find an item (tablet PC or digital camera) they would purchase given the opportunity.

Participants interacted with the system until they found the item they would buy. Participants were then given the opportunity to survey the entire catalog and to change their choice of item. Effectiveness was then measured by the fraction of participants who found a better item when comparing with the complete selection of alternatives in the database. So, using this metric, a low fraction represents high effectiveness. This method has also been used by researchers in marketing (Häubl and Trifts, 2000).

4.3.3 Find n-best items

In content-based recommender systems effectiveness has also been measured in terms of the keywords of recommended items. Symeonidis et al. (2008) evaluate explanations in terms of a metric they call "coverage". This metric is a weighted sum for the keywords that are used in an explanation, divided by the weighted sum for all the keywords that are important for a user. Thus, a good explanation uses many of the (distinguishing) keywords that are in a profile of a user[1]. For example, the coverage of the explanation mentioning the name of an actor considers if this actor occurs in the user's profile, and weighs this name against the (weights for all) the actors in this user's profile.

[1] The keywords are inferred from the movies that are rated highly by this user, which are both informative for this user and distinguish them from other users

In a similar manner, others have adjusted tradition accuracy metrics such as recall and precision (see also Section 2.3.1), and describe them in terms of the number of relevant (also weighted) keywords from a taxonomy in a user profile (Cramer et al., 2008a). Precision here is defined as the number of profile concepts matching with user interests, divided by the number of concepts in their profile. The evaluation in this case was centered around a task where participants were asked to select 6 favorite (best) items.

These types of metrics are heavily dependent on the recommendation accuracy of the recommender engine, as well as the choice and computation of important keywords. For example, if the recommendations are not correct the effectiveness of the explanations will be low as well. The fault to be corrected however, would lie in the recommendation engine. They also suffer from other weaknesses such as the granularity of keywords: more general keywords are likely to be right more often than more specific keywords, overlap between them (e.g. how to deal with near synonyms and word senses), as well requiring the availability of correct and rich meta-data (required not only for the items, but also for the effectiveness metric). These limitations w.r.t. to evaluation have been address by comparing several explanations in the same system (Cramer et al., 2008a).

These metrics validate the explanations for “accuracy”: whether the explanations contain the known interests of the user. They do not however, assess whether the user actually likes the item after trying it. This metric would not detect a fault in the user model that resulted in poor satisfaction with an item, but assuming good modeling this might be an inferred consequence (i.e. if we model and explain the user’s preferences well, they will like the item too). In comparison, the metric mentioned in 4.3.1 allows the computation of effectiveness for “bad” as well as “good” recommendations. That is, the metric is not dependent on the recommendation accuracy.

4.3.4 Our choice of metric of effectiveness

We elected to broaden the definition of effectiveness suggested by Bilgic and Mooney (2005), and described in Section 4.3.1. We recall that effectiveness can then be measured by the discrepancy between Steps 1 and 3 ($Rating1 - Rating2$) below:

1. **(Rating1)** The user rates the product on the basis of the explanation
2. The user tries the product (alternatively approximate by reading online reviews)
3. **(Rating2)** The user re-rates the product

The justification for our choice is that measuring the change of opinion allows us to compute effectiveness for a wide variety of items, including those that users initially evaluate to be uninteresting/poor. Since the important measurement here is the change of opinion, it is ok if the user does not like the item initially. If the user continues to dislike the item after trying it, this suggests that the explanation offered correct information. In contrast, if the user ends up liking the item after trying it, this would suggest a poor explanation.

We consider when it is appropriate to use the signed and absolute value for effectiveness. We use the *absolute* value of the difference between the two ratings when comparing different types of explanations to judge the general effectiveness, as well as the correlation between the two ratings.

Also, this metric does not give an indication of whether over- or underestimation is preferable to users, or if this preference might be a domain dependent factor. It also does not discuss whether an incorrect valuation of the same type (either over- or underestimation), but with different starting points, are comparable. When we want to see if there is more over or underestimation, we elect to study the signed values instead. We discuss the effect of domains and different starting points in Section 4.5.

4.4 Over- and underestimation of recommended items

User valuations of recommended items can be incorrect (over- or underestimations) or have poor effectiveness, for a number of reasons. For example, if the quality of the information used to form a recommendation, or if the recommendation accuracy is otherwise compromised, this is likely to lead to poor effectiveness. Also, the nature of the recommended item (e.g. relative cost) and presentation of the recommended items (see below) are likely to affect effectiveness. We discuss all of these factors below, but limit our exploratory studies to the effect of type of recommended object (in Section 4.5) and presentation of the recommended items (in Section 4.6) on *perceived* effectiveness.

4.4.1 Algorithmic accuracy

One reason the accuracy of recommendation may be damaged is that the recommendation algorithm is flawed. Another is that incorrectly rated (either over- or underestimated) recommendations may be due to insufficient information (e.g. low confidence recommendations), or a bias in data. Cosley et al. (2003) showed that manipulating a rating prediction can alter the user's valuation of a movie to cause either an over- or underestimation. For example, users (re-)rated movies (which they had rated previously) lower than their initial rating when they saw a lower prediction for the movie, and higher when the prediction was higher than their initial rating. The study also suggests that users can be influenced to change their rating of a movie from negative to positive. Cosley et al. (2003) does not discuss whether over- or underestimation is considered more severely by users, but did find that users' valuations of movies changed more often for lower predictions (underestimation) than for inflated predictions (overestimation). That is, users were more prone to under- than overestimation.

4.4.2 Presentational choices

Presentational choices for recommendations may also incorrectly influence a user's valuation of an item. For example it has been argued that order of presentation (Joachims et al., 2005), and the use of images (Nguyen and Masthoff, 2008) can have a persuasive effect on users. Joachims et al. (2005) found that users click more on highly ranked links, while Nguyen and Masthoff (2008) found that domain credible images could be used to increase credibility of websites.

4.4.3 Additional information

Assuming good algorithmic accuracy, additional information such as explanations can be used to either aid or hinder decision support. An explanation may contain both positive and negative information, and in that sense may have a polarity in a similar way to numerical ratings of a product. Modifying the polarity of an explanation is likely to lead to a similar influence on rating behavior as the one found by Cosley et al. (2003). Likewise, Bilgic and Mooney (2005) showed that some types of interfaces can cause overestimation (see also Section 4.2.3).

Online reviews are another form of additional information and might sway user valuation of an item. When we analysed the properties of helpful reviews, we found a positive bias in the movie domain (See Chapter 5). There were by far more positive reviews than negative, and positive reviews were considered more helpful by other users. Others have found a correlation between the helpfulness rating other users gave to a reviewer, and the rating this reviewer gave items in the domains of digital cameras and mobile phones (Kim et al., 2006).

In the experiment described in the next section, we study the effects of over- and underestimation (or damaged effectiveness) due to additional information such as explanations. However, since the over- or underestimation in the valuation of recommendations can be caused by any of these factors (e.g. limited algorithm, skewed or limited data, presentation, and additional information) the effects of evaluations of over- and underestimation may be generalized to these causes as well.

4.5 Experiment: Effectiveness in different domains

4.5.1 Over vs. underestimation

In this experiment, we wanted to find out whether users are more accepting of underestimation or overestimation in general. We also investigated how the nature of a product domain can mitigate, or conversely, exacerbate faulty information.

Domains

As we mentioned, the choice of product category is likely to affect the reaction that faulty information elicits in users. Our selection of product categories is motivated by previous work in the field of economics.

In economics, there has been a great deal of debate about classification of products into different categories. Shapiro (1983) uses the distinction between "experience goods", or goods that consumers learn about through experience, and "search goods" which they do not need to learn about through direct experience. Similarly, Cho et al. (2003) distinguishes between sensory products and non-sensory products. We propose an interpretation of these categories which distinguishes between products which are easy to evaluate objectively and those which commonly require an experiential and subjective judgment.

Another common categorization in economics involves investment or cost. Often this is a complex construct. For example, Murphy and Enis (1986) discusses *perceived* price in terms of the dimensions of risk and effort. This construct of risk includes financial risk but also psychological, physical, functional and social risk. The construct of effort considers purchase price, but also time that the purchase takes. Cho et al. (2003) also discuss perceived price in terms of non-monetary effort and degree of involvement. Laband (1991) narrows down the definition of cost to the objective measure of the purchase price of an item. For simplicity we will also use a definition of investment which only considers purchase price.

Given that item differ greatly in their nature, we choose to discuss products that vary at least along two dimensions: (financial) investment (or cheap vs. expensive) and search vs. experience goods (or subjective vs. objective domains).

4.5.2 Materials

The experiment was conducted using paper questionnaires. The questionnaires considered four domains distributed over the dimensions of investment (low vs. high) and valuation type (objective vs. subjective) as shown in Table 4.5 (see also Appendix A for examples of the questionnaires).

Table 4.5: Choice of domains

	Low investment	**High investment**
Objective	Light bulb	Camera
Subjective	Movie	Holiday

We defined investment in terms of price. By this definition cameras and holidays are high investment domains. Relative to these domains, light bulbs and movies can be considered low investment domains.

We considered cameras and light bulbs as objective domains, and movies and holidays as subjective. Our definition of this dimension is based on the premise that while some domains are highly subjective, it is easier to give a quantitative judgment in others. For example, users might be able to reach a consensus as to what properties are important in a camera, and what generally constitutes good quality, while this might be harder for a movie. It might be easier to define good image resolution in a camera than define good acting in a movie. Note also that our choice of definition for this dimension does not preclude that different product features (such as resolution and shutter speed, or actors and director) may vary in terms of importance to different users in all four product domains.

4.5.3 Hypotheses

We expect that users will be more lenient toward underestimation, and consider it more helpful than overestimation in general. This hypothesis is based on the assumption that users prefer being recommended only great items and missing ok ones (underestimation), to buying more, from being recommended items that they will not like (overestimation).

It also seems probable that users will have higher demands on accuracy in high investment domains such as movies and holidays. Likewise, users may respond more leniently to over- and underestimation in subjective compared to objective domains as these are harder to gage.

We also consider that it is possible that the strength of an over- or underestimation may also depend on the starting point on a scale. Therefore, we also consider the effects of over- and estimations of the same magnitude, but with different starting points. For example, what is the effect of underestimation on perceived effectiveness if a user's valuation of an item changes from negative to ok, and how does this compare to a change from ok to great? A user may consider an explanation least helpful when it causes them to perform an action they would not have performed if they had been given accurate information, e.g. when it changes their valuation of a product from good to bad, or from bad to good. Our hypotheses are thus:

- **H1:** Users will perceive information (such as explanations) leading to overestimation as less effective than (assumed explanations that cause) underestimation.
- **H2:** Users will perceive information (such as explanations) leading to over- and underestimation as less effective in high investment domains compared to low investment domains.
- **H3:** Users will perceive information (such as explanations) leading to over- and underestimation as less effective in objective compared to subjective domains.
- **H4:** Users will perceive information (such as explanations) leading to cross-over gaps, which cross the line from good to bad and vice-versa, as less effective compared to information resulting in other gap types.

4.5.4 Participants

Twenty participants (7 female, 12 male, one unspecified) were recruited at the University of Aberdeen. They were all postgraduates or researchers in Computing Science. The average age was 31.95 (range 20-62).

4.5.5 Design

We used a mixed design, with product domain as a within subject factor, and over- vs. underestimation as a between subject factor. Participants were assigned to one of two conditions. In the first, participants were given a questionnaire with overestimation scenarios, and in the second, underestimation scenarios (see also Appendix A for examples).

In the underestimation condition participants saw *Paragraph A*:

Paragraph A: *“Assume you are on a website looking for a particular product to buy (such as a camera, holiday, light bulb, movie). Based on the information given, you form an opinion of the product, and decide* ***not*** *to buy it and to spend the money on something else. Later you talk to a friend who used the product, and your opinion changes.”*

The user decides not to buy a product and spends the money on something else. This is to ensure that the choice (not to purchase) is perceived to be irreversible by the participants. Only later do they discover that the product was not as bad as they first thought.

For overestimation we considered situations in which the user initially rated the product highly, but then found the true value of the product lower after buying and trying it. *Paragraph A* is replaced with *Paragraph B* below:

Paragraph B: *“Assume you are on a website looking for a particular product to buy (such as a camera, holiday, light bulb, movie). Based on the information given, you form an opinion of the product, and decide to buy it. After using the product, your opinion changes.”*

In both cases participants were asked to consider that they were viewing a new website for each scenario even for similar products. All participants considered products in all four product domains (cameras, light bulbs, movies and holidays) in randomized order. Each participant was given scenarios in which their valuation of the product changed by a magnitude of 2 on a scale from 1 (bad) to 5 (good). We varied the starting point for the initial valuation. The rating of the product can be either:

1. Positive, i.e. staying on the positive side (3 ↔ 5)

2. Negative, i.e. staying on the negative side (1 ↔ 3)

3. Cross-over, i.e. changing from one side of the scale to the other (2 ↔ 4)

The order of the three starting points (positive, negative and cross-over) varied between participants in a latin square design. The orders of the before and after values were reversed between over- and underestimation, e.g. $3 \rightarrow 5$ (underestimation) became $5 \rightarrow 3$ (overestimation). Given three different starting points and four product domains, each participant considered twelve scenarios.

For each of the twelve scenarios, participants rated how helpful they found the (presumed) information given on the website on a seven point Likert scale (1 = very bad, 7 = very good): *“How do you rate the information on this website given this experience?”*. While this *perceived* effectiveness differs from true effectiveness, it also differs from persuasion. Persuasive information would give the user an initial impression (either positive or negative), but fails to consider the way the user finally rates the product once they try it. In this study the final rating is assumed to be known and true. Step 2 of the metric we have chosen to use (see Section 4.3.4), where the user would normally receive information about the product, is assumed to be a black box. In this sense, we use a stronger measure of perceived effectiveness than used by e.g. Hingston (2006) (see also Section 4.2.4).

Table 4.6: Perceived helpfulness (on a scale from 1 to 7) for over- and underestimation

	Mean (StD)	Median
Overestimation	2.59 (1.06)	2
Underestimation	3.08 (1.21)	3

Table 4.7: Perceived helpfulness (on a scale from 1 to 7) for the four domains

	Underest. - mean (StD)	**Overest. - mean (Std)**	**Underest. - median**	**Overest. - median**
Camera	2.87 (1.25)	2.37 (0.96)	3	2
Light bulb	3.15 (1.23)	2.63 (1.07)	3	2.5
Movie	3.30 (1.24)	3.00 (1.15)	3	3
Holiday	3.00 (1.14)	2.37 (1.00)	3	2

4.5.6 Results

Which is better?

Firstly we inquire if information leading to over- or underestimation is considered generally more helpful by users. Similarly we want to know just how harmful these over- and underestimations are considered by users. As can be expected, in Table 4.6 we see that information leading to both over- and underestimation are considered unhelpful. Since it is arguable that the values on a Likert scale may not be equal in distance, we performed a Mann-Whitney non-parametric test which rendered a significant result ($p < 0.01$). Overestimation is considered to be less effective than underestimation: H1 is confirmed.

Does the domain matter?

In Table 4.7 we offer an overview of perceived helpfulness, for all four domains.

Low vs. High Investment Table 4.8 summarizes the perceived helpfulness in low (light bulbs and movies) and high (cameras and holidays) investment domains. The perceived helpfulness was lower for high investment than for low investment domains (Mann-Whitney test, $p < 0.05$). A separate analysis for over- and underestimation shows a significant effect (Mann-Whitney test, $p < 0.05$ with Bonferroni correction) for overestimation, but not for underestimation. We also see that underestimation is considered as less effective in high investment compared to low investment domains, but this trend is not statistically significant. It seems as if users are more sensitive to information leading to over- and underestimations in high investment domains, but in particular to information leading to overestimation. H2 is confirmed.

Table 4.8: Perceived helpfulness for low vs. high investment domains

	Underest. - mean (StD)	Overest. - mean (Std)	Underest. - median	Overest. - median
High	2.93 (1.19)	2.37 (0.97)	3	2
Low	3.23 (1.23)	2.82 (1.11)	3	3

Table 4.9: Mean (and StD) of perceived helpfulness for objective vs. subjective domains

	Underest. - mean (StD)	Overest. - mean (Std)	Underest. - median	Overest. - median
Objective	3.00 (1.24)	2.50 (1.02)	3	2
Subjective	3.15 (1.19)	2.68 (1.11)	3	3

Objective vs. Subjective In Table 4.9 we see that information leading to both over and underestimation is considered less effective in objective compared to subjective domains, but the trend is not statistically significant. This hints that correct estimates may be more important in objective domains than subjective, regardless of direction of the error (over- and underestimation). User comments also confirm that some users are more forgiving of misleading information in subjective domains than objective: *"a wrong suggestion about 'subjective' evaluations of products (such as for movie or holidays) should not determine a severe bad judgment of the website."*, *"whether I like a movie (or holiday) is very subjective, and I would not blame my liking a movie less on the quality 1st description"*. The effect is however not sufficiently strong, and H3 is not confirmed.

Does the type of gap matter?

We hypothesized that information that leads to gaps which cross over between the positive and negative ends of the scale (cross-over gaps) will be considered less helpful than information that leads to the two other gap types. We found a significant effect of gap type on perceived effectiveness in a Kruskal-Wallis test ($p < 0.05$). However, in a Mann-Whitney test we found no significant difference between cross-over gaps and the two other gap types combined. H4 is not confirmed.

Table 4.10: Mean (and StD) of perceived helpfulness for different gap types

	Underest. - mean (StD)	Overest. - mean (Std)	Underest. - median	Overest. - median
Positive	3.90 (0.94)	3.02 (1.08)	4	3
Cross-over	3.03 (0.14)	2.68 (0.94)	3	2
Negative	2.31 (1.24)	2.05 (0.94)	2	2

Investigating the difference between gap types further, in Table 4.10 we see that participants would consider information leading to gaps on the negative end of the scale ($1 \leftrightarrow 3$) less helpful than gaps on the positive end ($3 \leftrightarrow 5$), and gaps which cross over between the positive and negative ends of the scale ($2 \leftrightarrow 4$), for data using both over and underestimation. Cross-gaps in turn were considered less helpful than positive gaps. For this reason we ran three post-hoc Mann-Whitney tests comparing the three gap types pairwise, all three were all found to be statistically significant ($p < 0.05$ with Bonferroni correction). Apparently, negative gaps damage the perceived helpfulness the most out of the three gap types rather than cross-over gaps.

A similar series of post-hoc Mann-Whitney tests were run for over and underestimation separately. All tests returned significant results (p < 0.05, with Bonferroni correction), except for the difference between positive and cross-over gaps for overestimation. That is, the difference in perceived effectiveness between positive and cross-over gaps for overestimation is negligible.

4.5.7 Discussion

Our finding of user preference for (information causing) underestimation compared to overestimation is in line with persuasive theory regarding expectancy violations and attitude change (Stiff, 1994). An audience's initial expectations will affect how persuasive they find a message. In a persuasive context, if expectations of what a source will say are disconfirmed, the message source can be judged to be less biased and more persuasive. For example, if a political candidate is expected to take a certain position with regard to an issue, but ends up advocating another position, their credibility rises.

Since it is a likely assumption that users expect a commercial recommender system to overestimate the value of an item, underestimation disconfirms this expectation and might cause users to find a recommender system less biased and more trustworthy. Two users stated expectations on an emphasis on high ratings in qualitative comments: *"I would expect the web to present items at their best and sometimes with some exaggeration."*, *"I expect there to be hype about a movie and to have to read between the lines to form a judgment for myself."*

The effect of gap type was surprising, we also were surprised to find that negative gaps were considered least helpful, and positive gaps most helpful, for *both* over and underestimation. This may reflect the way users distribute and assign ratings. The polar ratings of 1's and 5's are more uncommon and differently distributed from the other ratings, i.e. the 'distance' between 2 and 3 may be perceived as smaller than the distance between 2 and 1. So a user is much less likely to buy an item rated 1 rather than 2. Likewise, the probability of a user trying an item increases more between 4 and 5 than it does between 3 and 4. The lack of significant results for overestimation might be attributed to users' general expectation of overestimation in commercial recommender systems.

User comments also revealed some other interesting views on product categories. Two users left comments where they differentiate between holidays and the other products: *"Things like 'Holidays' matter more compared to goods, because holiday is a destination could be once in a life time thing."*, *"A holiday is an experience of value that cannot be replaced or compensated for, knowledge should be accurate."*. One user found it difficult to imagine using a recommender system to buy light bulbs: *"I can't imagine going on to a web site to look for information on a light bulb!"*.

4.5.8 Reflections on the experimental setup

When considering the design of our experiment, two criticisms can be raised. In this section, we discuss what these criticisms are, and why we decided to perform the experiment in this particular way.

Why the wording for underestimation differs

In the scenario for overestimation the user changes their value judgment by experiencing the product directly. In contrast, in the underestimation scenario, the user changes their value judgment based on comments from a friend who experienced the product. So, why did we not let the user "experience" the product directly in the latter case, as this would have made the conditions more comparable? As the user did not buy the product, it was hard to devise a plausible story of how they ended up experiencing it after all. If somebody else bought it for them as a gift, the user is not likely to regret missing the item, and thus will not harbor feelings of resentment over poor information to the same degree. Experiencing the item by borrowing it from a friend is not possible for all domains (e.g. holidays).

Why the experiment is indirect

Instead of participants really experiencing the products, we only told them about their experience. What participants think they would do in such a situation may diverge from what they really would do (Ajzen and Fishbein, 1977). We were however working on the basis of these assumptions:

- *Gap size matters.* Participants' perceived effectiveness will depend on the size of the discrepancy between their first impression and their valuation after experiencing the item.

- *Gap position matters.* The influence of an evaluation error (over- and underestimation) will depend on the gap's position. For example, an under-estimation from 1 (first rating) to 3 (final valuation) may have a different effect than one from 3 to 5. Evidence for this was found in our experiment.

Given these assumptions, for a fair comparison between domains (H2, H3) we need to control for gap size and position. Practically, this would mean that participant's valuations (before and after) need to be similarly distributed for all products. This would be hard to control rigorously. Even making the experiment a little more realistic, by giving participants particular information to form a first opinion, and then more information to form a final valuation, would be hard to control. Other researchers have failed to construct item descriptions with predictable ratings for all participants (Masthoff, 2004).

For a fair comparison between over- and underestimation (H1), we also need to control the gap size and position[2]. Suppose we knew that people *on average* like a particular item, and disliked another item. This may be hard to obtain in certain product domains, or limit us to a small subset of items where people converge on valuation. This is also likely to require a separate study to decide on suitable items. The estimated valuation would allow us to know, on average, the real valuation (and in analysis, we would need to remove all subjects whose valuation differed from this average). We would still have to make the explanations such that they induce the right initial rating (namely the valuation for the liked item in the disliked item's case, and the other way around). Given that we also wanted to study gap types (H4), we would need multiple of these item pairs plus explanations per domain.

Naturally, there are also strong advantages to running a similar direct experiment, in particular the feasibility that would be involved with a live system. A possible solution would be to run the experiment in a large scale system, and select data so with particular starting points and gap types (into bins), and then select equal sample sizes randomly out of each bin. As we mentioned however, this may not be feasible in all domains, the filtering can be said to be overly artificial, and the resulting data may involve a great deal of noise. It is certainly arguable which method would be most "correct". However, we maintain that in this case, an indirect experiment is justified. More direct (although not using a live recommender system) experiments are reported later in the thesis, as in Chapter 6.

4.6 Exploratory studies of presentational choices

A more detailed account of the questionnaire on explanations styles and the focus groups described below is also posted in an online appendix[3], but only summarized here for lack

[2] We consider the gap '1 to 3' to be comparable to the gap '3 to 1' w.r.t. to position

[3] http://sites.google.com/site/navatintarev/, posted Oct. 2009

of strong results.

We conducted a questionnaire comparing different explanation styles (collaborative, content-based and case-based), to see if they were perceived as more helpful, or more persuasive for the domains of movies and cameras. As such we were only measuring the effect of presentational choice on *perceived* effectiveness. In this exploratory study we found that:

- Explanations for movies were more persuasive and perceived as more helpful than explanations for cameras.
- The effect of domain, and the limited data do not allow us to draw any definite conclusions about which explanation style is most suitable in each domain.
- Perceived effectiveness was not significantly affected by whether explanations were positive or negative. Positive explanations were however more likely to influence users to buy products. This is the case for all three explanation types.

We also conducted seven focus groups totaling sixty-seven participants evaluating ten presentational interfaces. The main question of these focus groups was whether users prefer text or graphics. While we expected some variation between participants, we wanted to see if some interfaces had a majority of users preferring one medium over the other. As these preference may differ depending on the content of the explanations, we used a number of interfaces, and also aimed to receive qualitative feedback on these. While these focus groups did not study effectiveness, they were intended as a first step towards this aim. Our findings can be summarized as follows:

1. For a number of different types of explanation content users do not have a clear preference for text or graphics, but are more often confused by graphical interfaces.
2. Users vary in terms of which features they find important, and whether or not they want an image (such as of the leading actor/actress) to accompany explanations.
3. Users can be very critical of confidence and competence explanation interfaces.

We did not receive a conclusive answer as to whether participants would prefer text or graphics for any of the different types of explanation content, rather it would seem that the participants' opinions were divided. The qualitative comments for each set of interfaces have however been instructional.

4.7 Summary

We first surveyed the literature for evaluations where recommender system explanations have been evaluated for effectiveness. We describe metrics previously used for evaluating explanation effectiveness. We also investigated the effects that faulty decisions (over- and underestimation) can have in different domains, and for different points on a numerical scale in an experiment, on perceived effectiveness. We concluded the following:

- Information (such as explanations) leading to overestimation is considered less helpful by users than information that causes underestimation (H1).
- Information leading to overestimation is considered less helpful in high investment domains than in low investment domains (H2).
- Information leading to cross-gaps is not considered the least helpful by users, information leading to negative gaps is, for both over- and underestimation. For information leading to overestimation, positive gaps are not considered less helpful than cross-over gaps (H4).

As mentioned in Section 4.4, recommendations can be incorrectly biased for a variety of reasons. The results of this study would be relevant for algorithmic correction as well as studies comparing different presentational interfaces. Understanding the role of factors such as gap type, domain type and over and underestimation will help better control for these factors when optimizing a recommender system for effectiveness.

In light of our results we suggest an additional enhancement to the effectiveness metric proposed by Bilgic and Mooney (2005) and described in Section 4.3. We propose fine tuning this measure of effectiveness by weighting it according to gap type, over/underestimation and degree of investment.

We do not yet know what makes an explanation effective, and the type of information required is likely to be domain dependent. In the next chapter, we describe a number of studies in the movie domain which aim to answer this question, including corpus analysis and focus groups.

Chapter 5

Content for explanations: movies

5.1 Introduction

In this chapter we present the findings of several *exploratory* investigations aiming to identify the general properties of effective movie reviews, and discuss the possible implications of these findings on explanations in recommender systems. In this chapter we apply this methodology to the domain of movies. The studies discussed in this chapter consists of two analyses of online movie reviews retrieved from a commercial recommender system, two focus groups, and a two-part questionnaire. Most importantly, this chapter motivates the importance of personalization as we see that users differ in terms of what kind of information about a movie they consider important. A short list of relevant (domain specific) features is also suggested.

5.1.1 Methodology

The results outlined in this chapter are specific to movies, but the methodology used in this chapter can be used to elicit factors that are important for decision support in a large range of commercial domains. The analysis of reviews allow us to see what is mentioned in reviews, but also how the content reflects how helpful these reviews are perceived by users. Among other things this helps elicit which item features are commonly described. These findings can then be fortified by focus groups where we see if and how people describe these features in a more natural setting. Focus groups can also highlight additional factors in decision support beyond features, such as the relevance of the source of information. The formalization of item features used for decision support in that particular domain can also be fortified with more quantitative studies such as questionnaires.

5.1.2 Outline

We start with our analyses of online movie reviews. The first study of online reviews considered the properties of reviews rated as helpful by users, and reviews written by so-called top reviewers. In the second study, we investigated if these properties can be used to distinguish between helpful and unhelpful reviews. We also investigated which movie features were described in helpful reviews.

Helpfulness ratings of reviews give us an idea of how informative the content is, but do not give us an idea of how this information is used. We wanted to know if people described movies in a similar manner in natural conversation, as they do when writing reviews. While this was meant to be a more general exploration of what constitutes a helpful review, the main result is a list of commonly mentioned features (e.g. acting).

Among other things we wanted to see if and in what form these features reoccurred in natural conversation. In the focus groups we studied how people describe their favorite movies, and what factors and features participants used in the process of deciding about a movie.

Then, having a better insight into which features users may find important, we wanted to have a better idea of how many features it would make sense to mention and in what degree of detail. In a questionnaire, we investigated preferences for the number of features and degree of detail in manually constructed mock reviews. To the extent this was possible we controlled for confounding factors such as total number of words in a review.

5.2 Corpus analyses

The aim of analyzing movie reviews was to deduce general properties of explanations that were considered helpful by others. Reviews are a good example of evaluative judgment (Carenini and Moore, 2000b), and movie reviews in particular are amply available. The Amazon site [1] was chosen over other movie review sites for two reasons.

Firstly, the reviews justify purchase rather than other forms of consumption such as rental. The increase in degree of investment results in visibly more substantial reviews compared to e.g. MovieLens[2]. Reviews on Amazon not only reflect the polarity of users' evaluation of a movie (like/dislike), but also a more in-depth justification of their evaluation.

Secondly, reviews on the Amazon site are rated by other users as helpful or not - a function which may reflect what kind of reviews people like to receive, and not only what kind they like to write. The U.K. site was chosen to reflect local preferences and views.

[1] http://www.amazon.co.uk: retrieved November 2006

[2] http://movielens.umn.edu: retrieved November 2006

5.2.1 Study 1: properties of helpful reviews

This study aimed to identify the properties of helpful reviews. We were also curious if there was anything noteworthy about reviews written by Amazon's "Top-1000 reviewers" (in this chapter referred to as "top reviewers"). The criteria for reviewer ranking are not posted on the website. At a glance, a common factor is that these users have written a large number of reviews, although not all necessarily in the same domain. In other words, some of the top reviewers wrote reviews for domains other than movies.

To gain an initial intuition of which features are most commonly mentioned in reviews we informally browsed reviews on the Movie Lens recommender site. This progressed into a more structured analysis of fourty-eight reviews of DVD movies on the U.K. Amazon websites.

Methodology

In this study we surveyed reviews that were considered helpful to enough people, rather than valuable to a majority. A review is considered helpful if it is rated by at least five people, and out of these at least five consider the review helpful. Therefore, will still consider a review sufficiently helpful if fifteen peers rate the review, and five of them consider it helpful.

The selection of reviews was aimed not to bias any one movie or reviewer. There is therefore only one review for each reviewer where the reviewer name is known. Likewise there is only one review for each movie.

Amazon does not show more than ten reviews at a time and the reviews are by default sorted by recency. There is no ordering in terms of movie rating, unless explicitly requested by the user.

We chose to bias the selection of reviews toward negative reviews since the rating pattern of a user usually tends to be skewed toward positive ratings (Bilgic and Mooney, 2005). Thus, if there were negative reviews of movies fulfilling the criteria of helpfulness, these were given precedence over more positive reviews.

Movies were restricted to full length films, and excluded special editions, concert recordings, TV-series etc. Although they are not included in the analysis, we did note that reviews that describe aspects of a given DVD release, such as image or sound quality, were often considered helpful.

Results

Table 5.1 summarizes the most commonly mentioned features in online reviews. The number in brackets represents the number of reviews in which the feature was mentioned. The mean number of features mentioned was 4.5 (StD=1.65) across the corpus. Examples

Table 5.1: Most commonly mentioned features across forty-eight helpful online reviews. Numbers reflect number of reviews containing a feature (only counted once per review).

Cast (28)	Good in its genre (26)	Initial expectations (22)	Script (19)
Visuals (18) (incl. special effects, animations)	Suites mood (18)	Realistic (15)	Director (12)
Subject matter (12)	Easy viewing (8)	Good for kids (7)	Sex/violence (8)
Dialogs (6)	Pace (5)	Soundtrack (5)	Original (5)
Movie Studio (2)	-	-	-

Table 5.2: A comparison of top reviewers and other helpful reviews

	Top reviewer?	**N**	**Mean (StD)**
Summary length	Yes	29	142.10 (114.60)
	No	19	15.89 (34.61)
Total length	Yes	29	402.38 (264.62)
	No	19	196.95 (120.79)

of features can be found in Appendix B.3.

To check the reliability of the feature coding, a second coder annotated a 10% sample of the reviews. This sample consisted of 4 pairs of helpful (as rated by other users) and non-helpful reviews for a randomly selected set of movies, that is 8 reviews in total. The second coder suggested two new features: emotive response (e.g.“better than nothing”), and comparison with other movie (e.g.“there have been better films for that amount of money”). Emotive response was omitted from analysis, while comparison with other movie was reannotated as “previous expectations”. Using these parameters the agreement between the two coders was found to be 75%, and the inter-coder reliability was acceptable (Cohen’s Kappa = 0.34).

We also compared reviews written by top reviewers with those written by other “helpful” reviewers. In Table 5.2 we see that top reviewers tend to write longer reviews ($p < 0.01$), as well as write a longer summary compared to regular, but still helpful, reviewers ($p < 0.01$).

5.2.2 Study 2: helpful vs. non-helpful reviews

We conducted a second study to confirm that our initial findings could be generalized to differentiate between helpful and non-helpful reviews. In this study we analysed 74 user

Table 5.3: Most commonly mentioned features across thirty-seven pairs of helpful and non-helpful online reviews. Numbers reflect how many reviews contain the feature in the helpful/non-helpful versions.

Cast (20/20)	Good in its genre (20/12)	Initial expectations (2/1)	Script (18/13)
Visuals (16/8)	Suites mood (16/4)	Realistic (11/3)	Director (10/2)
Subject matter (10/9)	Easy viewing (8/0)	Good for kids (4/0)	Sex/violence (1/3)
Dialogs (6/1)	Pace (5/2)	Soundtrack (2/2)	Original (4/4)
Movie Studio (2/1)	-	-	-

reviews for 37 DVD movies, in a pair wise comparison of helpful and non-helpful reviews. We measured the same properties as in the first study, but also considered how understandable the reviews were. For this second study we used the Flesch-Kincaid Grade level (Kincaid et al., 1975), a known readability measure. The Flesch-Kincaid Grade Level score rates text on a U.S. grade-school level, so a score of 8.0 would imply that an eighth grader should be able to understand the document. We also measured the percentage of passive sentences, as these are considered to be decrease readability.

Results

Helpful reviews were longer ($p < 0.01$), and included a longer summary ($p < 0.01$), than non-helpful reviews. This finding is similar to the difference between helpful reviews and reviews written by top reviewers. The difference in length is also reflected in the fact that non-helpful reviews contain fewer features (mean=2.49, StD=1.45) than helpful reviews (mean=4.51, StD=1.69) for the same set of movies (based on first coder only). This difference is statistically significant ($p < 0.01$). Features that occurred much more frequently in helpful reviews were: director, suites mood, script, visuals, good in its genre, dialogs, easy viewing, good for children, realistic, sex/violence and pace. The features counts across the reviews are also summarized in Table 5.3.

In addition, we found that helpful reviews were more linguistically complex, with a higher Flesch-Kincaid Grade level ($p < 0.01$). The difference for the percentage of passive sentences was not significant however. The mean values are summarized in Table 5.4.

Discussion

We conclude this section on analyzing movie reviews with a note of limitation. Arguably, reviews are different from explanations in the context of a recommender system.

Table 5.4: A comparison of helpful and non-helpful reviews

	Average length (words)		**% Passive**	**Average grade level**
	Total	Summary		
Helpful	294.3	87.6	10.6	9.9
Non-Helpful	102	3.0	6.1	8.0

- Firstly, reviews can be both negative and positive, while it is arguable that explanations of recommendations are primarily positive. However, we found that negative reviews were by far rarer than positive ones. In addition, depending on the aim of the explanation, the recommender system may occasionally want to point out weaknesses of an item. For example, when aiming to increase overall effectiveness, explanations need to help users avoid poor choices as well as take advantage of good ones.

- Secondly, reviews often contain personal views, being more tailored to what the author finds important. Explanations on the other hand are more likely to be tailored to the user's set of priorities. Therefore, explanations will typically contain a subset of the available information, leaving out what is not important for a particular user.

- Thirdly, some types of explanations do not resemble reviews; for instance, they may explain the process of recommendation with the aim to improve transparency ("Others like you enjoyed this movie").

- Fourthly, users of a recommender system may also prefer short explanations than readers of reviews, or may not be able to read long descriptions for large numbers of items due to space limitations.

Despite these differences, an analysis of reviews aids our understanding of what makes helpful explanations, even though restricted to explanations that focus on item features. Using reviews rather than explanations integrated into a recommender system also allows us to survey textual properties in isolation without risking confounding with other properties of the overall recommender system.

5.3 Focus groups

The corpus analyses gave us an idea of which features users may consider when making decisions. However, the reviews reflect the interests of a speaker (or more correctly a writer), rather than a listener (reader). The helpfulness ratings work as a partial quality control, but focus groups allow us to delve deeper into the decision process, or how much these features determine whether or not a participants will like a movie. In particular, we aimed to find out how participants would like to be recommended, or dissuaded, from watching a movie.

5.3.1 Methodology

A total of eleven participants were spread over two focus groups, with the same facilitator. The discussion began with a short introduction of each participant's favourite movie. Later this progressed to a more detailed discussion for a set of movies. Participants were asked how long ago they saw the movie in order to insure they had watched the movie, and remembered their impression. They were asked about their initial expectations for the discussed movie, and if something in particular made them consider watching it. Participants were also asked about their impression after watching the movie, and what helped form this impression. Participants were additionally asked how they would like to be recommended or dissuaded from watching the discussed movie. If participants asked who was doing the recommendation/dissuasion, we told them this was someone who knew their taste in movies, such as a good friend.

We concluded with a summary of what had been said so far. The participants were asked for feedback on this summary. For completeness, they were also asked if any type of movie or deciding feature had been neglected in the discussion. At the end we went through the list of features mentioned in Table 5.1 and directly asked participants if they considered them when deciding whether or not to see a movie. In total, each session took between 1 and 1 1/2 hours.

Participants

Participants were recruited among the staff and student population of Aberdeen University, using fliers and poster boards. An interest in movies was mentioned as a prerequisite for participation. The sample is a self-selecting and voluntary group of people with a prior interest in cinema.

The participants consisted of eight males and three females, aged 24-33. With the exception of a French teacher, these were mainly staff or students of the computing science and maths departments. The participants also varied in nationality, with representatives from Ireland (1), Israel (1), France (3), Scotland (2), Spain (1), South Africa (1), Switzerland/Bolivia (1) and Vietnam (1). The fact that these participants come from an academic and multi-cultural workplace may lead to a bias in taste, such as an increased preference for foreign or independent cinema. We did, however, find a great divergence in the cinematic tastes of the participants, representing a wide spectrum of views.

Note that although numbering such as "Participant 1", is used in dialogs of multiple participants, this number is not consistent between dialogs and for the purpose of anonymity does not represent any one person.

Materials

The focus groups were audio recorded, transcribed and analyzed. Although the introduction served as a spring board of ideas of movies watched by the majority of the participants, we supplemented the focus groups with a pre-prepared list of movies (Appendix B.1.1). This list was based on the most commonly rated movies in the MovieLens 100.000 rating dataset [3]. Most movies are present in several genres, and we used a search by genre (most popular) on IMDB [4] to decide which genre they were best classified in. Occasionally, the genre annotation on MovieLens was clearly off mark. For example Star Wars was listed as a leading romantic movie, in cases like this IMDB was also consulted. The genres discussed included; action, children, animated, comedy, crime/gangster, documentary, horror, fantasy, musical, romance, science fiction, thriller and western.

5.3.2 Results

Introductions

Starting the focus groups with an introduction of each participant's favourite movies allows us to study which properties each participant intuitively considered important to mention. Citations of introductions are available in Appendix B.1.2.

The features mentioned varied between subjects; the most commonly mentioned feature was "good in its genre" followed by "script complexity" and "mood". Note that "mood" may relate to several sub-features in turn such as affect (Appendix B.1.2, quote 5), genre preferences (quote 8), and atmosphere (quote 11). Interestingly, two participants (in the same focus group) mentioned movies they enjoyed from childhood, which identifies a reason for watching movies (these introductions are annotated with "initial expectations" and "good for kids"). Table 5.5 gives a count of the mentioned features, and the number of times they were mentioned across all participants.

Modifications to features

Aside from a few modifications to the scope of each feature our focus groups largely confirmed the features deducted from the Amazon corpus. We first comment on the most

[3] http://movielens.umn.edu/: retrieved November 2006

[4] http://www.imdb.com: retrieved November 2006

Table 5.5: Number of times each feature was mentioned in participant introductions (focus groups)

Good in its genre (6)	Script (4)	Mood (4)	Subject matter (2)
Cast (2)	Initial expectations + Good for kids (2)	Director (1)	Visuals (1)
Realistic (1)	Original (1)	-	-

prominent modifications.

Firstly, we considered "realistic" to be a feature of a movie. Participants in both groups strongly differentiate between the terms realistic and believable. One participant explicitly stated: *"you used these two words and I think they are really important; realistic and believable. I don't care about it being realistic; I care about it being believable"*.

During the course of the focus groups, we also realized that *script complexity* was strongly tied to *mood*, although mood can be defined purely in terms of affect as well as in terms of genre. In addition, we realized that a simple script was interpreted synonymously with easy viewing. Some participants were happy to see movies as entertainment and did not place too much weight on the complexity of a story; others liked movies that presented a challenge, or were unpredictable: *"it depends a lot on how you come to the movies [Participant X] would like a movie that challenges him, do a bit of thinking. Personally, I pretty much think of a movie as a form of entertainment - two hours of fun!"*

There were also a number of new features that were mentioned when participants were explicitly probed about (potentially) additional features. Usually the suggestion was backed by a single participant. These included "gambling", i.e. watching a movie they didn't know anything about, although they might select the cinema: "Sometimes people just want to take a gamble. I've gone into movies totally not knowing what to expect - a total random pick". Another was the money a movie brought in on its opening night (although this participant did not find it decisive for themselves!). Participants did say that they could be affected differently depending on the types of awards a movie received such as Oscars and Cannes film awards: "Oscars put me off , Cannes turns on, sometime". Similarly to the comments about independent cinema, this was not perceived as a guarantee of a good movie, but a possible indicator of higher quality.

Inter-participant variation

Our focus groups confirmed that different users describe movies according to different criteria. In the dialogue below we see that Participant 1 is likely to differentiate movies according to director, while Participant 2 by the era in which the movie was filmed, and Participant 3 insists on the location in which the movie was filmed. The later two might both be clustered together into our initial category ”visuals“, but this dialogue also highlights the importance of differentiating between different types of visuals, and avoiding complete pigeon-holing:

Participant1: *“Did you prefer the Scarface by Hawks?”*
Participant2: *“The old one from the 50s, the black and white?”*
Participant1: *“Yeah, yeah.”*
Participant2: *“Yes, actually I do.”*
Participant 3: *“But that was set in Chicago, wasn’t it, that Scarface?”*
Participant2: *“I don’t remember the name of the town, but it was pretty much old fashioned.”*
Participant3: *“But it wasn’t the same Miami”*

In a similar manner particular participants were more aware of overall movie aesthetics and musical score while others did not notice or consider these features.

Mood

As mentioned earlier on, we realized that ’mood’ could describe several types of preference such as: script complexity, affect (e.g. feel good movie), as well as genre. These factors were often situational: *“I mean for a musical I don’t really need a great script, a great plot at least, uh or for uh what I call a pre-exam uh film the night before I mean. Bruce Willis saving the world is just what I need. Uh you know you don’t want something, you just want to use two neurons and that it, just relax.”*. However, they also depended on more constant factors such as genre preferences which varied between participants.

Social viewing

It became clear in both groups that for most participants there was a clear distinction between movies viewed in larger, more casual groups of friends, and movies seen alone or in more intimate circumstances such as with a partner. Movies seen with groups of friends were often light or easy viewing. Other movies, such as Schindler’s list were considered to be best viewed in more intimate company or even alone; *“I think I watched it on my own or something, I’m kind of thinking it’s not the kind of thing you watch at that age or in a group even.”*

The reason behind this seems two-fold. Firstly, participants felt that in larger gatherings the aim is often light-hearted entertainment, the viewers aim to enjoy themselves rather than conduct a mental activity. In contrast, they felt that more serious or dramatic movies may invoke strong emotions and tension and may be best viewed in more intimate company. Secondly, in large gatherings there is often a lot of simultaneous activity, someone is always speaking, going to get a tea or coffee etc. which may hinder the viewers from following a complex plot.

Who do you listen to?

Participants listened to their friends' recommendations, in particular when they had time to spare. Whether or not participants listen to a recommendation depends on how the recommendation was given: *"It probably depends on the way they describe the movie rather than who they are"*. Participants in both groups also agreed that the same advice coming from different people would not have the same impact on them. It depended on whether or not this person had similar taste, i.e. agreed on movies in the past: *"But it depends on the style of the movie; because if it's like a romantic comedy and my sister tells me its brilliant then I'll go and see it. If it's an action then I'll listen to what my brother thought of it. You know like different people like you know will, if I know they have similar tastes in that kind of film to me then I'll listen to them"*

Explanations and satisfaction

Reviews may help users enjoy movies more, rather than serve merely as decision aids. Participants believed that correcting faulty expectations for sequels or adaptations of a movie would not influence whether or not they saw it. Rather both focus groups unanimously felt that it could increase their acceptance upon viewing, and save potential disappointment. One participant stated that he liked musicals, but had to know what to expect in advance: *"If I go to see a musical I have to know it's a musical before watching it"*.

Initial expectations

Participants felt that reading a book tended to generate high expectations, e.g. *"I think the book was better, but I think I've never seen a movie that I've read the book that I've enjoyed the movie more. I think the book is always better."* They agreed that for adaptations of movies, expectations could be defused by information about how well the movie aligned with the original book. That is, in knowing that the film may diverge from the original on some counts, they would find it is "easier to forgive". As one participant said, *"that way the movie can be enjoyed in its own right"*.

Another factor contributing to expectations was the movie trailer. Sometimes the trailer was found to be better than the movie, which led to disappointment: "cus like sometimes with comedies the trailer gives you all the funny jokes so you go and those are ALL the funny jokes".

Dissuading users

None of the participants wanted to be completely dissuaded from watching movies they had disliked in the past. Participants even watched popular movies which they expected to be disappointed by. When asked if they would be upset with a friend for taking them to see a movie they did not like too much, participants replied they would not. During the course of both focus groups we gathered that participants wanted to form their own opinion, and they did not want to reject social invitations, or refute the general consensus without strong warrant: *"I wouldn't rush to watch certain genres, but if I was with somebody that was into that then yeah. I always think you try and take everything for what it is and try and look for the good parts"*. This reflects our findings of online movie reviews, and other datasets of ratings (e.g. Bilgic and Mooney, 2005). In our analysis of online reviews there were far fewer negative reviews than positive. Or, as one reviewer wrote: *"u cant just say a film isnt worth buying without any reasons!!"*

We suggest that this trend is mainly due to the social nature of movie viewing, and that social effects should be weaker for less social types of recommendations such as books or digital cameras. Movies can also be considered a low investment domain compared to others, which may explain why participants did not mind watching a movie they might be disappointed by (see also our discussion on the effect of domain on perceived effectiveness in Section 4.3). However, the social nature is likely to be the primary factor. One participant volunteered that he watched movies on his own, but that he would not suspend doubts about a particular movie if he was watching it alone *"I wouldn't go and see something I had doubts about, on my own"*. That is, the participant would be willing to suspend doubt in a social context, but the low investment in itself was not enough for him to do so.

What made you watch it?

In many cases, the participants watched movies due to social context, as mentioned above. A group of friends was going to the cinema, or were watching the movie at home. Another factor was availability, for example participants would watch a movie on television simply because they had a free evening, and a movie was on. A particular category may be reviewings of old favorites, such as the two participants who in their introduction described movies they enjoyed from childhood as favorites.

5.3.3 Discussion

In the corpus analysis, we had information about which reviews were considered helpful, while we do not have the same quantitative analysis of our focus group. As users may not truly know on what basis they form decisions this can be considered a limitation.

Nevertheless, in combination the corpus analyses and focus groups give us an idea of what kind of information users need to form their decisions about whether or not to watch a movie. The focus groups supply more detailed information, and allow users to express what is important for them in the decision making making process in a way that is not possible in reviews.

5.4 Features vs. Detail

In Section 5.2.1 we saw that top reviewers wrote longer reviews than other helpful reviewers, and in Section 5.2.2 that reviews that were considered helpful were longer than those considered not helpful. We also found that reviews with summaries were found to be more helpful. The summary in a movie review has no justifying purpose however, and we found that the participants in our focus groups often described movies without an intent to justify them. For example, we noticed that descriptive language was used to identify a particular movie, or differentiate it from another.

In a recommender system we are likely to be restricted in terms of the total length of an explanation, in particular when a user is exposed to many options at once. In this case, the explanations will have to strike a balance between the number of features (see Table 5.1 for a listing) and the detail in which they are described. In this pilot experiment we address two questions:

1. Do users prefer reviews with longer descriptions, or describing more features?
2. Is there any consensus on what features are important, and does this influence their preferences for Question 1?

5.4.1 Method

The experiment was a paper based questionnaire in two parts (see Appendices B.2 and B.3), each addressing one of the two questions described above.

The first part addresses the balance between longer and shorter descriptions. Here we ask participants to give their preference between two reviews, A and B, on a 7-point Likert Scale.

We created three reviews for comparison: Reviews I, II and III. Given the findings from our corpus analysis, we controlled the reviews for total length, the existence of a synopsis (identical in all reviews), and readability score. Table 5.6 summarizes the properties of these three reviews, see Appendix B.4. Given that recommender systems explanations are likely to be shorter than the presumed optimal length, we are consistent about this across all three reviews. Note also that Review I is necessarily longer than

Reviews II and III as it contains both more features and more detail. All reviews are linguistically complex (using the Flesch-Kincaid score described in Section 5.2.2), and well over the mean of non-useful reviews. Review III has an inevitably lower readability grade as lack of detail implies shorter (simpler) sentences.

The reviews were handcrafted, but based on a movie review (of the movie "The Constant Gardener") taken from our corpus analysis. We have removed all references to names, and personal evaluations such as '*'I thought the directing was brilliant"*. That is, the reviews can speak about directing, but not in first person (as in: "I thought..."). In our review with two features, we selected two features that were found informative in both the focus groups and analysis of reviews: genre and subject matter. The reviews with four features also described the director and visuals.

Table 5.6: Description of used reviews

	Words	**Synopsis**	**Flesh-Kincaid grade**
I. Detail 4 features	105	X	10.4
II. Detail 2 features	85	X	10.3
III. No-detail 4 features	77	X	9.0

We used a between subjects design with three conditions, with one comparison per participant. We randomized the order of the reviews A and B. The conditions are as follows (see Appendix B.4 for the review texts):

1. Condition 1: Detail, 4 features versus detail, 2 features
2. Condition 2: Detail, 4 features versus no detail, 4 features
3. Condition 3: No detail, 2 features versus detail, 2 features

As we had found in the focus groups that mood and scenario could be influential, we also supplied a scenario, which was consistent over all conditions. We chose the scenario to be as inclusive as possible: watching a movie with a close friend with similar tastes.

The second part was in place to see how strongly the preference for particular features influenced the answers in the first part. Here users were asked to tick up to five features they think are important for a review of any movie (the features were taken from the corpus analysis in Section 5.2, though participants were able to suggest other features). The rationale for this second part of the questionnaire is to control for particular feature bias. For example, participants who are more interested in director than they are in genre may prefer four features (a bias for Review III) due to this fact rather than the number of features.

We also saw this as an opportunity to see how much consensus there is about importance of features, increasing empirical strength while avoiding any rater/personal bias our results may have suffered from in our previous investigations.

Hypotheses

Our hypotheses are that:

- **Hypothesis one**: Condition 1, Review I vs. II . Amount of detail being fixed, users prefer more features over less.
- **Hypothesis two**: Condition 2, Review I vs. III. Number of features being fixed, users prefer more detail over less.
- **Hypothesis three**: Condition 3, Review II vs. III. If forced to choose, users will prefer more details and fewer features to less details and more features.

Participants

Thirty-eight computing students of the University of Aberdeen participated in the experiment. Of these 7 were female, 26 male, and 5 are unknown. The ages of participants range from 17 to 56, with a mean of 25.9. One of the questionnaires was incorrectly filled and was discarded.

5.4.2 Results

Part one: A or B?

We can see the cross-tabulation of votes in Table 5.7. The results of the this study did not render significant results (chi-square, assuming uniform distribution), as our participants did not have a clear preference for either A or B in either of the three conditions. The trends for each condition are summarized below.

- Condition 1: Amount of detail being fixed; participants showed a weak preference for more features over fewer. In line with hypothesis 1.
- Condition 2: No clear trend - number of features being fixed; some participants prefer more detail and others less. Not in line with hypothesis 2.
- Condition 3: Participants prefer more details and fewer features, to less detail and more features. In line with hypothesis 3.

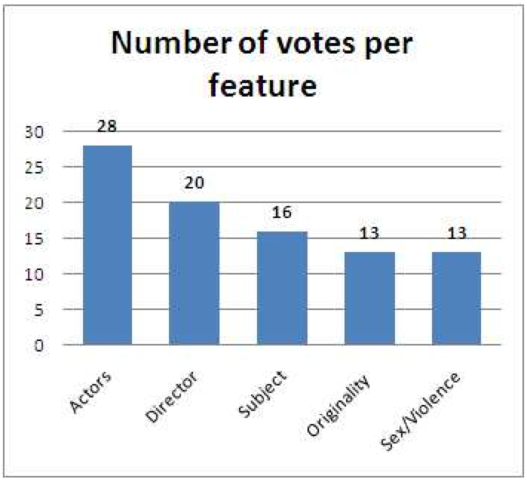
Table 5.7: Cross tabulation of preferences across conditions (count)

		Condition		
Likert rating		1: detail, 4 feat. vs. detail, 2 feat.	2: detail, 4 feat. vs. no detail, 4 feat.	3: no detail, 4 feat. vs. detail, 2 feat.
More A	1	0	1	3
	2	6	1	3
	3	2	4	2
	4	1	0	1
	5	1	1	0
	6	2	1	1
More B	7	2	3	2
Total		14	11	12

Part 2: Which features?

No feature received less than 5 votes. We survey the five most popular features in Figure 5.1. We note that the feature “director” received 20 votes. Participant comments suggest that this is a positive preference, i.e. director can persuade rather than dissuade, which may have led to participant preference being skewed in the direction of four features over two in the first part of the study. This may have influenced the trend in Condition 1, however, in the third condition (which forces participants to choose between more detail or more features), more participants chose detail even though this means missing out information about the director.

Figure 5.1: Number of votes for the leading five features



5.4.3 Discussion

Our questionnaire allowed participants to leave qualitative comments. From the comments left we see that several of the participants had a preference for brevity. What one participant found *"professional"*, another found *"too arty"* or considered additional names as "name dropping". A participant in Condition 3 who preferred the version with more features and less details explained that he preferred simpler reviews: *"Because it's more straight to the point and brief. Reviews shouldn't be too lengthy."*. That is, even when controlling for word length, detail can be perceived as verbosity.

We also see that users differ in their degree of interest in features and details, described in terms of "opinions" and "plot" respectively. Although the summary was exactly the same in all reviews, participants were sensitive to the relative proportions in both directions: *"Because there is more opinions about the film rather than going on and on about what happens in the film"; "describes the plot more than how filmed."*

Conclusion

Unfortunately the results of the experiment were not conclusive. Despite this, we have gained several new insights:

- Participants varied in their preference in the balance between the amount of detail and number of features.

- For some users balance seems to be calculated not only in terms of word counts, but also in relative proportions such as between summary and length of opinion.

- If forced to choose, participants might be more likely to prefer fewer features, but described in more detail. We hasten to caution that this result was a trend rather than statistically significant.

In hindsight, we consider if the three reviews were too similar as we controlled for lexical choice, length and Flesch-Kincaid grade level. The (lack of significant) results may an artifact of this particular review set. This might be remedied in a larger scale study with reviews with different topics and features. Another possibility, would be to consider individual differences in a within subjects design. In this case, each participant would conduct comparisons in all three conditions, giving us a relative comparison. For the purpose of this thesis we have however elected not to pursue this line of research further.

5.4.4 Guidelines for recommender systems

In the following sections we summarize the findings of our studies in terms of possible implications for explanations in recommender systems. Although some of our findings are merely trends, we hope that these initial studies may serve as a starting point for heuristics that can be used in any, not only movie, recommender systems.

Length

As we mentioned in Chapter 3, previous work has inquired whether concise explanations could be more persuasive and trust inducing respectively (Carenini and Moore, 2000a; Chen and Pu, 2002). In our corpus of online movie reviews, we saw that longer reviews are generally considered more helpful. We believe that there is a rough optimal range, perhaps somewhere between 100 and 200 words (see Tables 2 and 3). However, this does beg the question of whether explanations can be as long as reviews. Likewise, this does not say whether these reviews are truly effective, merely that they are perceived to be so.

Detail vs. number of features

It seems there is no consistent preference for reviews with more details or more features, but the trend suggests that when users are forced to choose they may prefer to have fewer features described in more detail.

Presence of a summary

The presence of a summary appears to have a positive effect in the context of movie reviews. It is possible that in other evaluative domains, a descriptive paragraph without a purely justifying purpose could have the same positive effect.

The aim of the explanations in a recommender system may define the degree of importance for this type of descriptive paragraph. For example, it may be more important for a recommender system aiming at higher satisfaction rather than persuasion - more descriptive explanations may increase user satisfaction with the system over all, while omitting the paragraph may function better in terms of getting the user to try an item.

Context

We found that users view different movies in groups than alone, and that mood may affect pace, script complexity or genre. To cater for this diversity, a recommender system should be susceptible to variations in context. We suggest that a recommender system be able to allow a user to specify their current priorities, and save a number of such profiles.

According to the findings of our focus groups, a recommender system aimed at groups as well as individuals would require knowledge of all the users involved including their willingness to disclose strong emotions, shared interests (for subject matter), or the general aim of the evening (e.g. pure entertainment).

Personalization

By taking into consideration which properties are important for each user it may be possible to cut down the length of the explanation (compared to reviews). In our studies we saw that users weigh features differently, even if there is a general consensus about which features are generally important. In the corpus analysis and focus groups we also saw that it is often more important to refer to the general quality, such as the general level of casting (e.g. good, bad) rather than mention particular actors. Naturally, in a recommender system users that have particular preferences should be able to set up filters accordingly.

Linguistic considerations

We see that movie reviews tend to have a high linguistic level, measured by the Flesch-Kincaid Grade Level, with more complex reviews often being considered more useful. Another possible restriction may also be lexical choice, it may be important that the words used reflect those commonly used in the domain. In several pilot studies of automatically generated reports for literacy learners (Williams and Reiter, 2008) found that the users had a significant preference for the less common, longer word "correct" over the more common "right" (in the spoken British National Corpus [5]). As the corpus of tutor written reports did not contain the word "right", (Williams and Reiter, 2008) suggest that this word should not have been used in reports.

Features

As an initial heuristic we suggest that an explanation presents around 4 features, since the mean in our movie review corpus lies at 4.5 features per review. Depending on the scenario, such as if the movie is seen in a group, different features may be more important. Also, participants in the focus groups believed that their mood is likely to initially influence the genre they choose to see, and as a secondary effect, what factors they consider important. Subject matter and how realistic a movie is are very relevant for a documentary or historical movie. In genres such as Action and Science Fiction, realism seemed to be watered down to a much less important feature of "believable" which is important in the negative sense, e.g. flaws in coherence. These results were consistent between participants, we would recommend that these findings be used for default settings. A user

[5] http://www.natcorp.ox.ac.uk/, retrieved Oct. 27, 2008

should be able to modify their preferences in greater detail if the generalization does not apply to them, however.

In more general terms, aside from the general plot - which was assumed following the analysis of online reviews - the participants in our focus groups explicitly stated that the script complexity and dialogues were most important, and genre was mentioned very frequently. As previously mentioned, importance for each particular feature varies between users, and should be tailored.

Final remarks

The remarks from the focus groups such as; *"But it depends on the style of the movie; because if it's like a romantic comedy and my sister tells me its brilliant then I'll go and see it. If it's an action then I'll listen to what my brother thought of it. You know like different people like you know will, if I know they have similar tastes in that kind of film to me then I'll listen to them."*; suggest that personalized explanations can be based on a collaborative algorithm as well. For example, we envision explanations of the type: *"User X likes the same type of thrillers you do, such as 'Silence of the Lamb's. User X liked 'The Usual Suspects' too."* A content based algorithm (for a user that likes information about genre and actors) on the other hand could result in explanations of the type: *"This movie is a very funny comedy staring Jim Carrey ! "*

5.5 Summary

In this chapter we describe a methodology that can be used in a multitude of commercial domains, to elicit which features are used to decide whether or not to try the item in question:

1. Analyses of online reviews to find out what features users consider important when making decisions.
2. Focus groups to see which features users use to describe items they have tried in the past, and in particular which way these features are referred to.
3. Questionnaires to quantify the important of features and answer any additional questions.

In addition, this chapter motivates the importance of personalization as we see that users differ in terms of what kind of information about a movie they consider important. As we also found that longer reviews were generally considered more helpful, we propose that some of this length can be cut down using personalization, taking advantage of the

Table 5.8: Final list of features

Rank	**Corpus**	**Focus groups**	**Questionnaire**
1	Cast	Genre	Cast
2	Genre	Script	Director
3	Expectations	Mood	Subject
4	Script	Subject	Originality
5	Mood	Cast	Sex/Violence

fact that users find different features most important. As a deliverable of the methodology we presented a list of movie features that users see as important, as well as a number of heuristics for explanations in recommender systems. Table 5.8 summarizes the leading features in all three studies. The features that repeat are cast, genre, subject, script and mood. In the next chapter, we describe an experiment in which we consider some of the movie features elicited, and study the effect of personalization on effectiveness of explanations.

Chapter 6

Personalization experiments

6.1 Introduction

The aim of this chapter is to consider how user-tailoring of item features can affect explanation effectiveness, persuasion and user satisfaction. While similar, our work differs from the studies in Carenini and Moore (2001) and Herlocker et al. (2000) (described in Chapter 4), which primarily considered the *persuasive* power of arguments and explanations, but did not study effectiveness. Arguably Carenini and Moore (2001) varied the polarity (i.e. good vs. bad) of the evaluative arguments, but given the domain (real-estate) it was difficult for them to consider the final valuation of the item, i.e. whether the user would really like the house once they bought it. Bilgic and Mooney (2005) suggested a metric for effectiveness, but did not consider the role of user-tailoring.

We have conducted user studies to elicit what helps users make decisions about whether or not to watch movies (see Chapter 5). We then used the elicited item features in a testbed natural language generation system, using commercial meta-data, to dynamically generate explanations (see Appendix C for more information about the implementation).

We conducted experiments in order to inquire whether personalization of the generated explanations helps increase effectiveness and satisfaction compared to persuasion. The general experimental design is described in Section 6.2. The first experiment (described in Section 6.4) gave surprising results: non-personalized explanations were more effective than personalized, while personalized explanations led to significantly higher satisfaction. Baseline explanations however, also did surprisingly well. For this reason, the experiment was repeated with stricter control for possible confounding factors. The results and modifications are described in Section 6.5.

As the results of these two experiments were surprisingly, we wanted to investigate whether the results were due to the domain, or if they could be generalized to a second domain. For this purpose we repeated the experiment in a more objective and higher investment domain: cameras. We describe this third experiment in Section 6.6. Surprisingly, also in this domain we found that participants made better decisions in the non-personalized condition, but preferred the personalized explanations. We summarize and suggest justifications for our results in Section 6.7.

6.2 General experimental design

All three experiments follow a similar design:

1. Participants were told that they were going to be asked to rate randomly selected items and their explanations which can be both positive and negative. We included negative explanations because we saw in our focus groups that while users did not want to be dissuaded from watching a movie, they did want to know at what level to set their expectations[1].

2. Participants rated the importance of different features and entered their preferences, resulting in a simple user model.

3. Participants evaluated a number of recommendations and explanations for items selected at random from a pre-selected set. Note that the explanations tell the user what they might think about the item, rather than how the item was selected. Moreover, these explanations differ from explanations of recommendations as they may be negative, positive, or neutral - the point is to help users to make good decisions even if this leads them not to try the item. For each item:

   (a) Participants were shown the item and explanation, and rated on a 7-point Likert scale (from bad to good):

   - *How much they would like this item.*
   - *How good the explanation was.*

   They could opt out by saying they had "no opinion", and could give qualitative comments to justify their response.

   (b) Participants read user and expert reviews on Amazon, care was taken to differentiate between our explanation facility and Amazon. This step serves as an approximation of actually trying and evaluating the item.

[1] To avoid overly biasing participants' ratings of the items, the explanations did not explicitly suggest a positive or negative bias such as in: "Watch movie A because ...", or "Do *NOT* buy camera B because ...". We wanted the information to be such that allowed the participants to freely make their decision without any unnecessary persuasive or dissuasive power. As such, one could argue that the information being given to participants was a description rather than an explanation. However, we argue that explanations may be descriptive if their role is to help decision support rather than to persuade or dissuade.

(c) They re-rated the item, and the explanation.

Persuasion can be seen as the initial rating for an item. Dissimilarly to effectiveness, this metric *disregards* the user's second rating *after* trying the item. While the user might initially be satisfied or dissatisfied, their opinion may change after exposure. Effectiveness is measured using the metric described in Chapter 4, and considers how the user's valuation of the item changes. Satisfaction is measured through the rating of the explanations.

In a between subjects design, participants were assigned to one of three degrees of personalization:

1. **Baseline:** The explanation is neither personalized, nor describes item features.
2. **Non-personalized, feature based:** The explanation describes item features, but the features are not tailored to the user.
3. **Personalized, feature based:** The explanation describes item features, and tailors them to the user's interests.

We hypothesize that:

- **H1:** Personalized feature based explanations will be more effective than non-personalized and baseline explanations.
- **H2:** Users will be more satisfied with personalized explanations compared to non-personalized and baseline explanations.

6.3 Limitations on the used explanations

The reader will note that the explanations used in these experiments are short, and rather simple. There are a number of reasons for this. Firstly, brevity is important in a context where the user has to review many possible options. An explanation plays a different role from e.g. a review that is more complete but requires much more time to read. Secondly, the features that are currently available (or will be available in the near future) in existing commercial services are limited in both diversity and depth[2]. It is harder to correctly extract certain features (e.g. what kind of mood a movie is suitable for) and understand them well (e.g. if this is a particularly strong role by a given actor or actress) than to extract other simpler features (e.g. the names of actors or directors). There may be natural language processing techniques that can be adapted for these cases, but this would

[2] In these experiments we have chosen to use Amazon Webservices as a representative example, although similar limitations are likely to occur with other commercial services.

require a deviation from the main focus of this thesis. Thirdly, assuming that complex features such as those mentioned above can be deduced about movies, it is much more difficult to create an algorithm that can infer user interest in these features. On the other hand, algorithms which consider simple features such as actor and director names already exist (e.g. Symeonidis et al., 2008). The previous work however has not considered how selecting which features to mention affects the effectiveness of explanations.

Thus, the question we are investigating here is whether the explanations that can be created with item meta-data (which already exists in a representative commercial system) can be made more effective through personalization. Or, in other words, if it would make sense for the developers of explanations in recommender systems to change their algorithms to explain by using item features, and if it makes sense to consider which item features to present in the explanation for a given user. We are aware that even a simplistic change in this direction is likely to be a large investment, and so an “offline” experiment of this type would be of great value before any implementation in an existing system. An alternative to this which would be of lower cost for developers, would be to use existing algorithms, but to change the *explanations* to use personalized item features. This is however a “deception” (w.r.t. to transparency of the recommendation algorithm) of users, and would not make much sense unless this helped the users to make better decisions. Let us now see if this is the case.

6.4 Experiment 1: Movies I

The aims of the initial experiment in the movie domain was to see if using movie features (e.g. lead actors/actresses), and personalization in explanations could affect their effectiveness and satisfaction for users. We also wanted to know how personalization affected persuasion: if we help users make decisions that are good for them (effectiveness), will they end up buying/trying fewer items (persuasion)?

6.4.1 Materials

Fifty-nine movies were pre-selected as potential recommendations to participants. Thirty are present in the top 100 list in the Internet Movie Database (IMDB [3]) and the other twenty-nine were selected at random, but all were present in both the MovieLens 100.000 ratings dataset [4] and Amazon.com. The requirement that half of the movies be present in the top 100 in the IMDB follows from the baseline condition described in our design below.

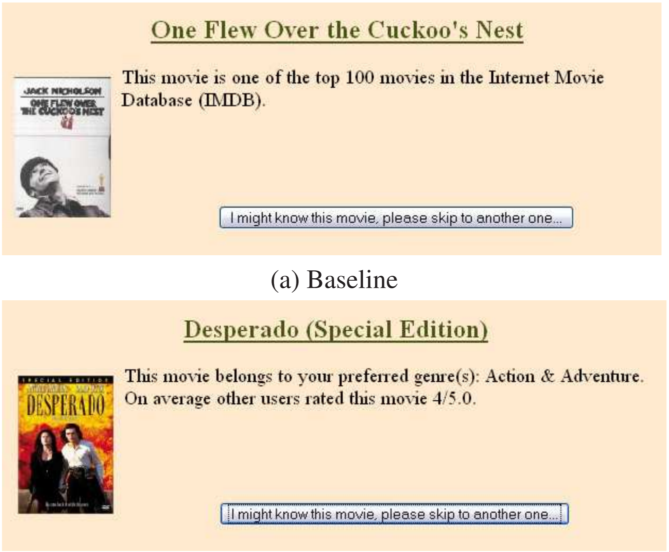
We chose to use item features that realistically could be extracted from a real world

[3] http://www.imdb.com

[4] http://www.grouplens.org/node/12#attachments

system such as Amazon web-services as these are freely available via an API for a number of domains, while considering the features extracted from our user studies in Chapter 5. This resulted in the features: *genre, cast, director, MPAA rating (e.g. rated R) and average rating*. Director and cast were previously found to be important, and MPAA rating reflects the two polar ends of “good for kids” and “sex/violence”. The average rating of the movie in reviews, or popularity, was not found to be important in any of studies (unless it can be seen as a by-product of being “good in its genre”), but was included as it was available.

6.4.2 Design



One Flew Over the Cuckoo's Nest

This movie is one of the top 100 movies in the Internet Movie Database (IMDB).

I might know this movie, please skip to another one...

(a) Baseline

Desperado (Special Edition)

This movie belongs to your preferred genre(s): Action & Adventure. On average other users rated this movie 4/5.0.

I might know this movie, please skip to another one...

(b) Random choice feature - the average rating is not necessarily selected as the user’s most preferred feature.

Figure 6.1: Example explanations, Movies I

First, participants entered their movie preferences: which genres they were in the mood for, which they would not like to see, how important they found other movie features, and the names of their favourite actors/directors. The user model in our testbed can weigh the movies’ features according to feature utility, and considers each participant’s genre preferences. See Appendix B.5 for screenshots of how the the user model was obtained.

Each participant evaluated *ten* recommendations and explanations for movies selected at random from the pre-selected set. Participants were assigned to one of three degrees of personalization (see Figure 6.1 for screenshots):

1. **Baseline:** *“This movie is one of the top 100 movies in the Internet Movie Database (IMDB).”* or *“This movie is not one of the top 100 movies in the Internet Movie Database (IMDB).”*
2. **Random choice, feature based:** The explanation describes the genres a movie belongs to and a movie feature. The movie feature mentioned is selected at random,

e.g. *"This movie belongs to your preferred genre(s): Action & Adventure. On average other users rated this movie 4/5.0"*. The feature 'average rating' may not be particularly important to the user.

3. **Personalized choice, feature based:** The explanation describes the genres a movie belongs to and a movie feature. The explanation describes the one item feature that is most important to the participant (rated the highest), e.g. *"Although this movie does not belong to any of your preferred genre(s), it belongs to the genre(s): Documentary. This movie stars Ben Kingsley, Ralph Fiennes and Liam Neeson your favorite actor(s)"*. For this user, the most important feature is leading actors.

Our user studies suggest that genre information is important to most if not all users, so both the second and third condition contain a sentence regarding the genre in a personalized way. This sentence notes that the movie belongs to some of the user's disliked genres, preferred genres, or lists the genres it belongs to though they are neither disliked nor preferred. Also, a movie may star one of the user's favorite actors or director in which case this will also be mentioned as a *"favorite"*, e.g. "This movie starts Ben Kingsley, Ralph Fiennes and Liam Neeson your *favorite* actor(s)."

If a participant had previous knowledge of the movie, they could request a new one by clicking on a button that said "I might know this movie, please skip to another one".

6.4.3 Results and discussion

Participants

Fifty-one students and university staff participated in the experiment. Of these, five were removed based on users' comments suggesting that they had either rated movies for which they had a pre-existing opinion, or Amazon's reviews instead of our explanations. Of the remaining, 25 were male, 21 female and the average age was 26.5. Participants were roughly equally distributed among the three conditions (14, 17 and 15 respectively).

Enough to form an opinion?

Table 6.1: Opt-outs (%)

Condition	**Movie Before**	**Movie After**	**Expl. Before**	**Expl. After**
Baseline	8.8%	0%	2.2%	0%
Random choice	7.2%	3.6%	3.0%	3.0%
Personalized	3.1%	0.6%	0.6%	0%

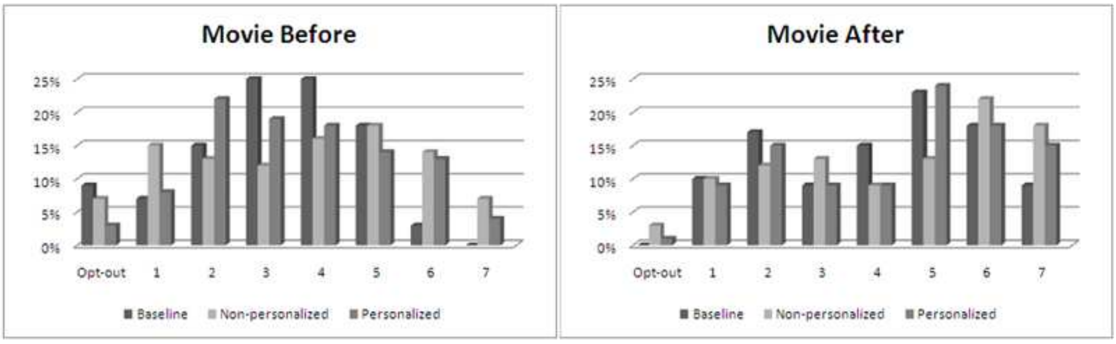
Since our explanations are very short we first considered whether they were sufficient for the user to form an opinion of the movie. If the explanations were too short to form any opinion of the recommended item, there would be little point in continuing the analysis. In Table 6.1 we note the percentage of no-opinions in each condition. We see that this is small though not negligible. The percentage for the first movie as well as for the first explanation is smallest in the personalized condition.

For our explanations to be relevant to participants, it is also important that the ratings of the movies vary. That is, participants are not just saying that every item is “ok”, selecting 4 which is in the middle of the scale, but are able to form opinions that are both positive and negative. In Figure 6.2 we consider the actual ratings of the movies. Here we see that the first and second rating of the movie are distributed beyond the mean rating of 4, suggesting that participants are able to form polarized opinions. We note however that a larger percentage of ratings in the baseline condition revolve around the middle of the scale compared to the other conditions.

Table 6.2: Means of the two movie ratings (excluding opt-outs) and mean of effectiveness between conditions. “Before” and “After” denote the two movie ratings before and after viewing Amazon reviews.

Condition	**Movie Before**	**Movie After**	**Effectiveness (absolute)**	**Effectiveness (signed)**
Baseline	3.45 (1.26)	4.11 (1.85)	1.38 (1.20)	-0.69
Random choice	3.85 (1.87)	4.43 (2.02)	1.14 (1.30)	-0.57
Personalized	3.61 (1.65)	4.37 (1.93)	1.40 (1.20)	-0.77

Figure 6.2: First and second movie ratings - the distribution is considered with regard to the *percentage* of ratings in each condition.



Are Personalized Explanations More Effective? (H1)

Next, we considered effectiveness. Table 6.2 summarizes the means of the movie rating when seeing the explanation (Movie Before) and after reading the online reviews (Movie

Table 6.3: Effectiveness over absolute values with "no-opinions" omitted, and Pearson's correlations between the two movie ratings.

Condition	Correlation	p
Baseline	0.43	**0.00**
Random choice	0.65	**0.00**
Personalized	0.58	**0.00**

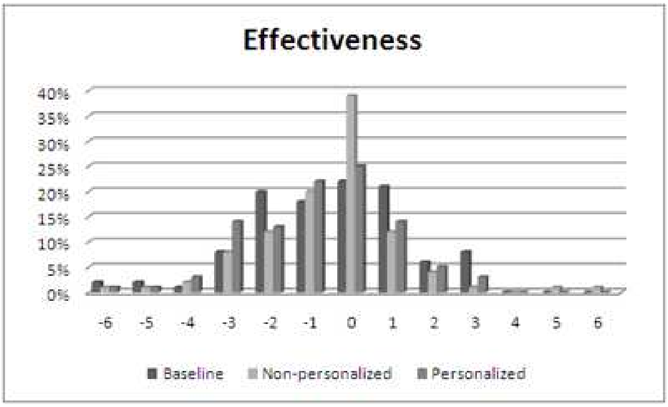
After). It also describes the mean effectiveness in each condition, using the *absolute* and unsigned value of the difference between the two movie ratings.

Similar to the metric described by Bilgic and Mooney (2005) we consider the mean of the difference between the two movie ratings. Unlike Bilgic and Mooney (2005) (who considered the signed values) we consider the *absolute*, or unsigned, difference between the two ratings in Table 6.3. We compare the difference across all trials, and participants, per condition: *a Kruskal-Wallis test shows no significant difference between the three conditions w.r.t. to effectiveness.* This suggests that the degree of personalization or using item features does not increase explanation effectiveness.

Figure 6.3 graphically depicts the *signed* distribution of effectiveness. We see here that underestimation is more frequent than overestimation in all three conditions. We also note the peak at zero in the random choice, feature based condition. Around 40% of explanations in this condition are perfectly effective, i.e. the difference between the two ratings is zero. A Kruskal-Wallis test comparing between all three conditions did not show significant differences in the initial ratings. We also looked at factors such as the relative proportion of the shown movies that were in the top-100 on IMDB for the two conditions, and the distribution of preferred features per condition, but did not find anything that would lead us to believe that there was more overestimation in the random choice condition.

Since Bilgic and Mooney (2005) did not consider the sign of the difference between

Figure 6.3: Distribution of (signed) effectiveness - "no opinions" omitted



the two ratings, their metric of effectiveness also requires that the two ratings are correlated. This correlation is still interesting for our purposes. Table 6.3 shows a significant and positive correlation between these two ratings for all three conditions. *That is, explanations in all three conditions perform surprisingly well.*

Are users more satisfied with personalized explanations? (H2)

In Table 6.4 we see that the second set of explanation ratings are higher than the first. This may be partly due to some participants confounding our explanations with the Amazon reviews, thus rating our explanation facility higher for Explanation After. For this reason, we do not compare the Explanation Before and After ratings. The mean rating for Explanation Before is low overall, but a Kruskal-Wallis test shows a difference in ratings between the conditions. Mann-Whitney tests show that users rate the first explanation rating significantly highest in the personalized condition ($p<0.01$). This suggests that while the personalized explanations may not help users make better decisions, users may still be more satisfied. This is confirmed by the qualitative comments. For example participants in the personalized condition appreciated when their preferred feature was mentioned: *"...explanation lists main stars, which attracts me a little to watch the movie..."*, while they felt that vital information was missing in the random choice condition: *"...I indicated that Stanley Kubrick is one of my favorite directors in one of the initial menus but the explanation didn't tell me he directed this."*

Table 6.4: Means of the two explanation ratings (excluding opt-outs) in the three conditions.

Condition	**Explanation Before**	**Explanation After**
Baseline	2.38 (1.54)	2.85 (1.85)
Random choice	2.50 (1.62)	2.66 (1.89)
Personalized	3.09 (1.70)	3.14 (1.99)

6.4.4 Summary: Movies I

In all three conditions participants largely have an opinion of the movie, and in all conditions there was more underestimation than overestimation. The mean effectiveness deviated ca 1.5 from the optimum discrepancy of zero on a 7 point scale (StD < 1.5), regardless of the degree of personalization or whether or not the explanation used features such as actors. In light of this under-estimation we reconsider the fact that movie ratings in general, and their Amazon reviews in particular, tend to lean toward positive ratings. If Amazon reviews are overly positive, this may have affected our results.

Since there is no significant difference between conditions w.r.t. effectiveness we consider the factors that the three conditions share, which is that they all expose the participant to the movie title and movie cover. A number of participants justify their ratings in terms of the image in their qualitative comments, in particular for the baseline explanation. So it is fair to assume that at least some participants use the image to form their judgment.

We note however that the correlation between before and after ratings is significant in all three conditions, and strongest in the random choice condition. However, participants were significantly more satisfied with the personalized explanations. The strong result for the baseline was surprising[5].

6.5 Experiment 2: Movies II

In this section we repeat the experimental design of the initial experiment, taking into consideration a number of possibly confounding factors.

6.5.1 Modifications

Before rerunning the experiment we applied a number of modifications. Firstly, in the previous experiment the cover image was displayed in all three conditions, which may have helped participants make decisions. In order to investigate the influence of cover image on effectiveness, this experiment presents explanations *without* images.

Secondly, we ensured that information about genres is more detailed and complete in this experiment, as participants complained that genre information automatically retrieved from Amazon was incorrect and incomplete. The genre information was annotated by hand, and the generated explanations describe all the genres a movie belongs to.

Thirdly, the differentiation between the personalized and random choice explanations was not sufficiently distinct in the previous experiment, and has been made more clear in this follow-up. The random choice condition describes all the genres of the movie, but no longer relates them to the user's preferences - making this condition truly non-personalized[6]. Also, the random selection of feature in this condition previously considered *all* the features, but now excludes the one rated the highest by the participant.

Fourthly, the previous experiment typically took participants around 45 minutes to complete. These participants may have been fatigued by the end of the experiment. For this reason, we reduced the number of trials from ten to three.

[5] It is arguable that this may be because participants were trying to be consistent between the two ratings. We note however that we did not preselect any option or display the first rating when asking the users for the second rating, and that we required participants to read reviews inbetween the two ratings. Therefore, remembering the initial rating and purposeful consistency would have required an additional effort on the behalf of the participants.

[6] Likewise, when the movie contains the user's favorite actor, we not only mention this actor, but all of the leading actors.

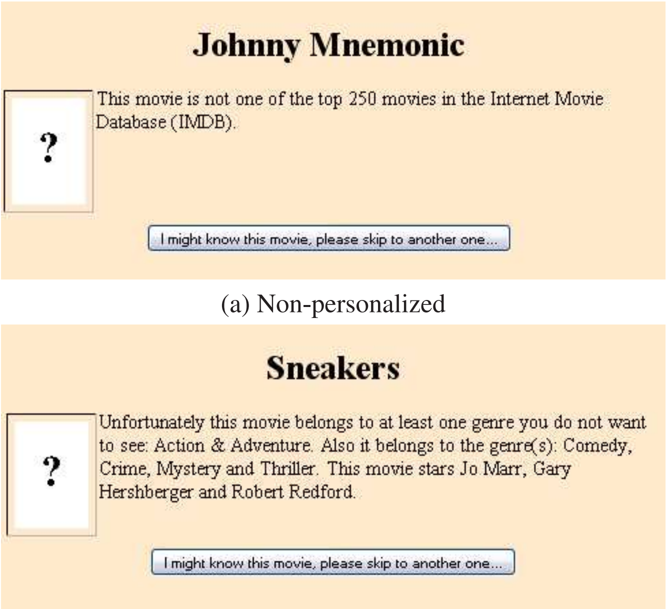
Finally, a minor modification was made in the baseline condition, to mention movies in the top 250 (rather than 100) in the Internet Movie Database (IMDB).

6.5.2 Materials

Eighty-five movies were pre-selected as potential recommendations. The movies were distributed evenly among 17 genres. As a movie belongs to multiple genres, they were balanced according to the main genre. Fourteen of the movies were present in the top 250 movies in the Internet Movie Database (IMDB).

Movies were also selected for having a high degree of variation of rating. High variation is more likely to lead to polarized views leading to an even distribution of initial ratings of movies. We used the measure of rating variation (entropy) described in Rashid et al. (2002), based on the MovieLens 100.000 ratings dataset [7].

6.5.3 Design



Johnny Mnemonic

This movie is not one of the top 250 movies in the Internet Movie Database (IMDB).

?
I might know this movie, please skip to another one...

(a) Non-personalized

Sneakers

Unfortunately this movie belongs to at least one genre you do not want to see: Action & Adventure. Also it belongs to the genre(s): Comedy, Crime, Mystery and Thriller. This movie stars Jo Marr, Gary Hershberger and Robert Redford.

?
I might know this movie, please skip to another one...

(b) Personalized

Figure 6.4: Example explanations, Movies II

We repeated the design of the first experiment, considering the modifications described in the previous section. Participants were assigned to one of three degrees of personalization (see Figure 6.4 for screenshots):

1. **Baseline:** The explanation is neither personalized, nor describes item features: *"This movie is (not) one of the top 250 in the Internet Movie Database (IMDB)".*

[7] http://www.grouplens.org/node/12#attachments

2. **Non-personalized, feature based:** e.g. *"This movie belongs to the genre(s): Drama. Kasi Lemmons directed this movie."* The feature 'director' was not particularly important to this participant.

3. **Personalized, feature based:** e.g. *"Unfortunately this movie belongs to at least one genre you do not want to see: Action & Adventure. Also it belongs to the genre(s): Comedy, Crime, Mystery and Thriller. This movie stars Jo Marr, Gary Hershberger and Robert Redford."* For this user, the most important feature is leading actors, and the explanation considers that the user does not like action and adventure movies.

6.5.4 Results and discussion

Participants

Fourty-four computing science students participated as part of an HCI practical. Eleven were removed from the analysis: six did not complete the experiment, and five completed it in under 90 seconds. Of the remaining thirty-three, twenty-six were male and seven female. The average age was 24.58 (StD=6.58). Participants were roughly equally distributed between the personalized and non-personalized conditions (16 and 11 respectively). We note that a majority (seven) of the participants that were removed from the analysis were in the baseline condition. It would seem probable that participants in this condition felt discouraged by the lack of complexity in these explanations. This condition has therefore been removed from further analysis, save for qualitative observations.

Enough to form an opinion?

Table 6.5: Opt-outs (%)

Condition	Movie Before	Movie After	Expl. Before	Expl. After
Baseline	55.6	16.7	44.4	16.7
Non-pers.	4.3	**0**	**0**	**0**
Pers.	15.2	15.2	3.0	9.1

Table 6.5 shows the percentage of opt-outs per condition. We see that the proportion of opt-outs before reading reviews is least for the non-personalized condition (for both movie and explanation ratings). Likewise, the proportion of opt-outs after reading reviews is zero in the non-personalized condition (for both movie and explanation ratings). In contrast, there are more opt-outs for the movie rating given before reading reviews in the baseline (10/18 trials) and personalized condition (5/33 trials).

Are personalized explanations more effective? (H1)

Table 6.6: Means (StD) of movie ratings (excluding opt-outs) and effectiveness in two conditions

Condition	Movie Before	Movie After	Effectiveness (absolute)	Effectiveness (signed)
Non-pers.	3.84 (1.95)	3.93 (1.95)	0.96 (0.81)	-0.09 (1.25)
Pers.	3.75 (2.05)	4.00 (1.87)	1.33 (1.27)	-0.25 (1.85)

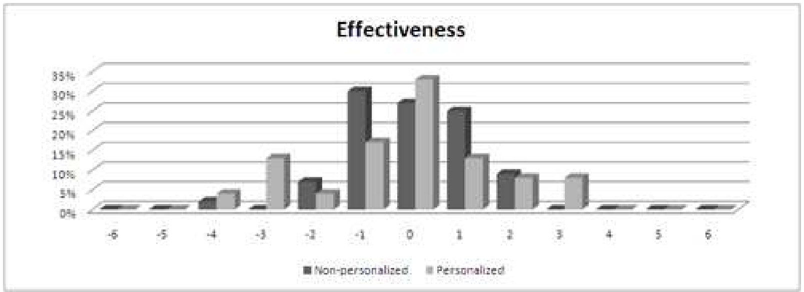
Table 6.6 shows the means of the movie rating when seeing the explanation (Movie Before) and after reading the online reviews (Movie After). It also shows the mean effectiveness in each condition, using the *absolute* value of the difference between the two movie ratings. As we are trying to minimize the gap between the two movie ratings, good effectiveness is denoted by smaller values. We see that the mean of the "Movie Before" and "Movie After" ratings are roughly equivalent for both conditions. Since the initial item ratings are similar, the explanations are comparable in terms of persuasion. *Effectiveness appears to be best in the non-personalized condition, but as in the first experiment this difference is not significant (Mann-Whitney test).*

Figure 6.5 shows the distribution of effectiveness across the two conditions. Here we use the signed difference in order to be able to discuss the direction of skew. Table 6.6 shows that in all conditions there is a slight tendency toward underestimation. [8]

Table 6.7: Effectiveness correlations between before and after item ratings

Condition	Correlation	p
Non-pers.	0.79	**0.00**
Pers.	0.56	**0.01**

Figure 6.5: Distribution of (signed) effectiveness per condition (opt-out ratings omitted)



We note also that the mean values for effectiveness reported here are comparable to those in the first experiment (see Tables 6.3, 6.6 and 6.7). Since this experiment uses

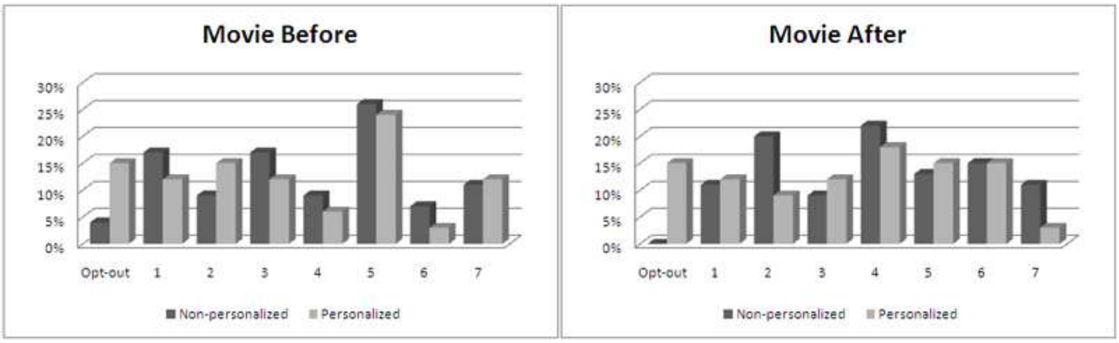
[8] This may be due to using an e-commerce website to gain additional information, which in fact causes an overestimation in the final valuation.

different movies and a different number of trials per user, they can not be compared statistically. However, the results still indicate the reliability of results over two experiments, and that the presence or absence of cover image for movies is not likely to greatly affect effectiveness compared to other factors for feature-based explanations.

We also consider the correlation between the first and second movie rating. We reiterate that for this measure to be relevant, the ratings should be diverse. For example, it is possible to imagine a selection of movies that are considered ok on average, in which case the before and after ratings are both 4 (on a scale from 1 to 7). This should not be the case given the selection criteria for movies.

Figure 6.6 shows the distribution of movie ratings, both before and after reading online reviews. Table 6.7 shows the Pearson correlations between the two movie ratings per condition (a similar result is found if Spearman's correlation coefficient is used.). The strongest correlation is in the non-personalized condition, and significant but weaker in the personalized condition ($p < 0.01$ for both). These findings are contrary to our hypothesis that personalized explanations would be more effective.

Figure 6.6: Distribution of movie ratings



Are users more satisfied with personalized explanations? (H2)

Table 6.8 shows the average values for explanation ratings. We see that the mean for personalized explanations is highest (Explanation Before[9]). This coincides with the results of our first experiment, but the difference between the non-personalized and personalized conditions is not statistically significant this time.

Qualitative comments

We consider if there is some effect of title on effectiveness as we chose not to omit it. A single participant (non-personalized condition) commented that they made use of the title to make a decision: *"Mostly I based my decision on description of the movie (the movie*

[9] Similarly to our first experiment in the movie domain, we do not include a comparison with Explanation After ratings as these may reflect participants' valuation of Amazon rather than our explanations.

Table 6.8: Means of the two explanation ratings (excluding opt-outs) in the two conditions.

Condition	Explanation Before	Explanation After
Non-personalized	2.72 (1.68)	2.83 (1.74)
Personalized	3.31 (1.55)	2.97 (1.33)

belongs to my favourite genres), title sounds quite interesting too." Some participants did not see the baseline as an explanation at all: *"Explanation says nothing";"still no explanation"*.

6.5.5 Summary: Movies II

We found explanations in the non-personalized feature based condition to be significantly more effective than personalized explanations. It is also in the non-personalized condition that participants opted out the least. Within the personalized condition, the movie rating before reading online reviews and after were however still strongly and significantly correlated. The personalized explanations themselves were given higher initial ratings, but this time the trend was not statistically significant. That is, in both experiments we found that non-personalized explanations were better for decision support than personalized, despite the fact that our participants were more satisfied with personalized explanations.

The significant difference between the personalized and non-personalized conditions compared to the previous experiment is likely to be due to addressing potentially confounding factors in the previous experiment: making a clearer distinction between the personalized and non-personalized condition, and ensuring complete and correct genre information. The removal of a cover image did not seem to damage effectiveness notably in any of the conditions. However, we did find that many participants in the baseline condition either opted out of giving ratings, clicked through the experiment, or dropped out altogether. This suggests that explanations such as our baseline without images could damage user satisfaction considerably, although not necessarily decision support.

We offer a few possible justifications why non-personalized explanations did better than personalized in these two first experiments. Firstly, users may not know how to best represent their preferences. It has been shown before for example that implicit ratings may result in more accurate recommendations than explicit (O'Sullivan et al., 2004).

Secondly, the features as we represented them are rather simplistic and a more sophisticated representation could be more effective. E.g. it may be useful to say if this is a strong role by a particular actor, or to use the actor's face rather than name (as we saw in focus groups on presentational interfaces, see also Section 4.6) . We were curious to find out if this is a property of a subjective domain, where it is difficult to quantify such features. In the next section we describe a similar experiment in a more objective domain (cameras), to see how this affects our results.

6.6 Experiment 3: Cameras

6.6.1 Introduction

In the previous sections we found that non-personalized explanations were better for effectiveness, while personalized led to higher user satisfaction, in the movie domain. This was a surprisingly result and we wanted to investigate whether the results were due to the domain, or if they could be generalized to a second domain. For this purpose we repeated the experiment in a more objective and higher investment domain. Our intuition is that the movie *domain* suffers from being subjective in nature. So while it is possible to talk about the user's favorite actor starring in a film, the actor's performance may be deemed as both good and bad depending on the user. Nor is an actor's performance likely to be consistent across their career, deeming this feature (most commonly selected by our participants) a poor indicator for decision support. We would expect this effect to be smaller in a more objective domain such as digital cameras.

We have also seen that participants may be less forgiving of overestimation (persuasion) in high investment and (relatively) objective domains (see Section 4.5). We are interested to see how additional and personalized information in explanations influences users in the camera domain: whether this impacts effectiveness, persuasion, and user satisfaction.

6.6.2 Features for cameras

Due to time constraints we did a less rigorous analysis of important features than we did for movies. We did however want to elicit which features are generally considered important when purchasing a camera. As an initial guideline, we surveyed which features existing recommender systems in the camera domain have used (Chen and Pu, 2007; Felfernig et al., 2008; McCarthy et al., 2004, 2005). We shortlisted the following features: brand, optical zoom, price, resolution, weight, memory and type of camera (SLR or point and shoot). From these memory was excluded as modern cameras usually have external memory that can be added on. The remaining six features are all readily available on Amazon's webservice.

Next, in a questionnaire 11 members of staff at university or members of the university photography club (1 female, 10 male; average age 44.67, range 29-62) rated the importance of these six features. A copy of the questionnaire can be found in Appendix B.6.1. The purpose of the questionnaire was twofold. Firstly, we wanted to know whether there are features that are commonly considered important. Secondly, we wanted to find out if there was a case for personalization, i.e. do different people find different features important. Overall, the type of camera, brand, and price were found to be most important. However, this was not a complete consensus, people do rate the features differently. It is not the case that any one feature is rated highest by each participant, leaving some scope for personalization.

6.6.3 Modifications

Firstly, users are less likely to be consumers of cameras than movies. To exclude participants that would never use or buy a camera, participants indicated their photography expertise and likelihood of purchase.

Secondly, the non-personalized explanation describes the three most commonly selected item features to a user, which were camera type, brand, and price. Three features are mentioned to make these explanations comparable in length to the explanations in the movie domain which mentioned genre as well as a feature. Likewise, the explanation in the personalized condition describes the three item features that are *most* important to a user.

Thirdly, there is no strong equivalent to the Internet Movie Database for cameras, so the baseline needed to be replaced. We chose an interface which is similar to the explanations given on a number of commercial sites: a bar chart which summarizes review ratings of the camera categorized into good, ok, and bad. This is an explanation available on the Amazon website. It is similar to the barchart used in Herlocker et al. (2000), but considers all reviews rather than similar users only.

It is imaginable that it would be possible use the ratings of similar users in the context of a collaborative-filtering algorithm. The underlying data for this algorithm could not be based on review ratings however, as the similarity computation would suffer from sparsity and low overlap between items (see also sparsity in collaborative-filtering in Section 2.2.4). Amazon likely remedies the problem of sparsity by using purchase data and/or viewing patterns rather than reviews. Also, a bar chart of similar users for each item would require a fully operational recommender system which each participant would need to interact with extensively. This is not necessary, and possibly not even desirable, for our experimental purposes.

6.6.4 Materials

People buy less cameras than they see movies. This means that participants are unlikely to be familiar with a particular camera, especially because explanations were accompanied with a generic image and hard to identify. For this reason we did not need a "I might

Table 6.9: Range, means, and counts of features over the 22 cameras used in this experiment.

Feature	Range	Mean (StD)	Mode
Price	106-1695£	448.40 (489.23)	225.73
Resolution	5-12 megapixels	9.45 (2.21)	10
Zoom	1-10 x	5.77 (4.80)	3
Weight	118-930 g	421.59 (286.86)	334
Camera ‘type’	SLR (n=9), ”point-and-shoot“ (n=13)		
Brands	Panasonic (n=4), Nikon (n=4), Canon (n=4), Olympus (n=4), Fujifilm (n=3), Sony (n=3)		

Table 6.10: Total number of ratings and mean number of reviews per item (StD), by category of rating

Good		Ok		Bad	
Total n	Mean (StD)	Total n	Mean (StD)	Total n	Mean (StD)
326	16.09 (13.10)	19	0.86 (1.22)	25	1.32 (2.69)

know this item” button as in the movie experiments, and only used a small dataset.

Twenty-two cameras have been hand-picked as potential recommendations. Specifications for SLR cameras were defined by the lens that came with them per default. Table 6.9 summarizes the range for each of the features. It is also possible to select the cameras automatically via an API, but handpicking the items enabled us to control the range for each feature better. See also Appendix C for details about implementation.

As seen in Table 6.10, there were by far more good ratings (4’s and 5’s) than ok (3’s) and bad (1’s and 2’s), which is a pre-existing bias for the cameras on the Amazon website which had at least 3 reviews[10].

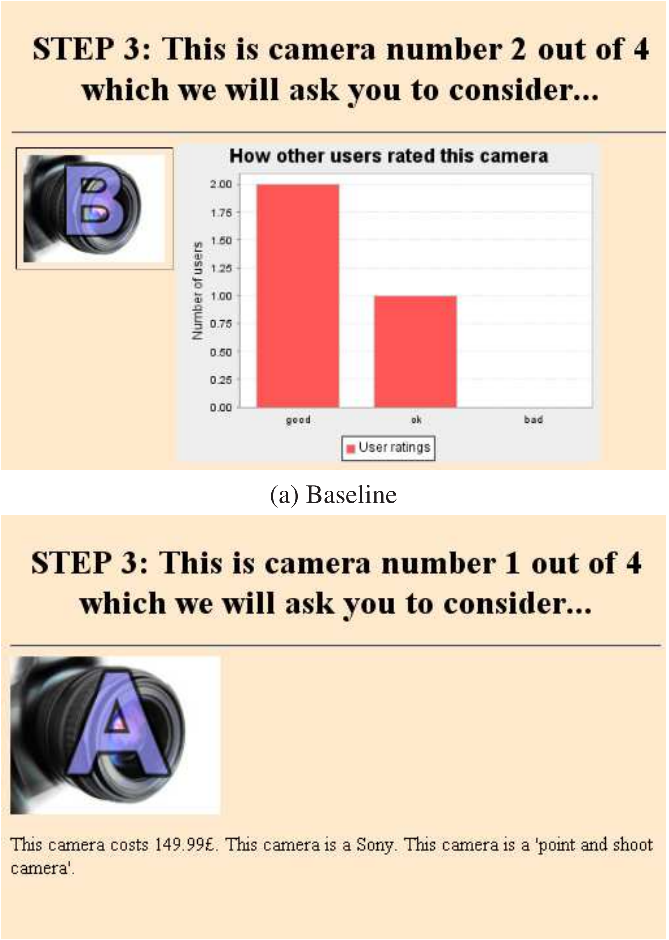
6.6.5 Design

Participants evaluated 4 recommendations and explanations in total. First, they specified information about themselves, including their self-perception of their photography expertise, and likelihood of purchasing a camera. Next, they rated their preferences for the features (see Appendix B.6.2 for screen shots.)

Participants were assigned to one of the three degrees of personalization of the explanations:

1. **Baseline:** This was a bar chart such as Figure 6.7 a.
2. **Non-personalized:** e.g. *“This camera costs 179.0£. This camera is a Panasonic.*

[10] This was a minimal requirement for selection of the cameras, as the baseline explanations which were based on the review ratings would have been uninformative otherwise.



STEP 3: This is camera number 2 out of 4 which we will ask you to consider...

(a) Baseline

STEP 3: This is camera number 1 out of 4 which we will ask you to consider...

This camera costs 149.99£. This camera is a Sony. This camera is a 'point and shoot camera'.

(b) Non-personalized

Figure 6.7: Example explanations

This camera is a ’point and shoot camera’.” (See Figure 6.7 b).

3. **Personalized:** E.g. If features ’price’, ’brand’ and ’zoom’ are most important the explanation may be: *“This camera costs 679.95£. This camera is a Nikon. It has an optical zoom of 11.0x.”*

In all three conditions, the explanation was accompanied with an identical image of a camera. Between cameras the only difference in the image was that a semitransparent letter (A-D) was superimposed over the image to differentiate the four cameras.

6.6.6 Revised hypotheses

The first two hypotheses, H1 and H2, are the same as before (see Section 6.2), but the changed baseline results in a third hypothesis. The reviews for cameras are strongly biased toward positive ratings: there are more positive reviews than negative and neutral. Bilgic and Mooney found that a positively biased bar chart is likely to lead to overestimation of items (Bilgic and Mooney, 2005). For this reason we also hypothesize that:

- H3: Users are more likely to overestimate their rating of the camera in the baseline condition compared to the two feature-based explanations (persuasion).

6.6.7 Results

Participants

Fifty-two students and university staff participated in the experiment. Five were removed from analysis: one for not completing the experiment, three for being "unlikely to buy a camera" and one for saying they "knew nothing about photography".

This left forty-seven participants, distributed between the three conditions: baseline (n=17), feature-based non-personalized (n=15) and personalized (n=15). The average age was 24.17 (StD=5.85) with a range of 18-48. Thirty-one participants were male and sixteen female.

Enough to form an opinion?

Table 6.11: Opt-outs (%)

Condition	**Camera Before**	**Camera After**	**Expl. Before**
Baseline	23.9%	7.5%	6%
Non-personalized	16.7%	8.3%	3.3%
Personalized	1.6%	0%	3.2%

As in the previous experiments we inquire if the short explanations are sufficient for users to form an opinion. Table 6.11 shows that there was a larger percentage (23.9%) of opt-outs for the first camera rating in the baseline condition compared to the other conditions.

Table 6.12: Means of the two camera ratings (excluding opt-outs) and effectiveness per condition.

Condition	**Camera Before**	**Camera After**	**Effectiveness** *(absolute value)*	**Effectiveness** *(signed value)*
Baseline	3.94 (1.47)	4.75 (1.73)	1.77 (1.50)	-0.77 (2.20)
Non-personalized	3.88 (1.62)	4.78 (1.75)	1.14 (1.32)	-0.78 (1.57)
Personalized	3.83 (1.86)	4.95 (1.77)	1.88 (1.34)	-1.08 (2.05)

Are personalized explanations more effective?

Table 6.12 shows the ratings of the cameras and effectiveness per condition. Comparing between conditions we found a difference in (absolute) effectiveness (Kruskal-Wallis,

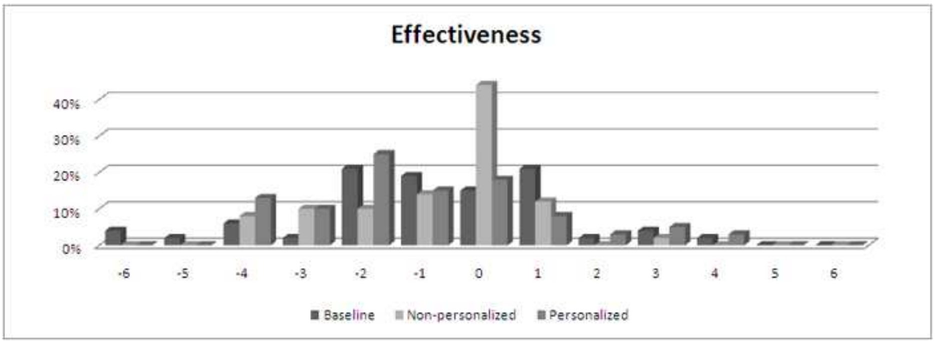


Figure 6.8: Distribution of (signed) effectiveness per condition (opt-outs omitted)

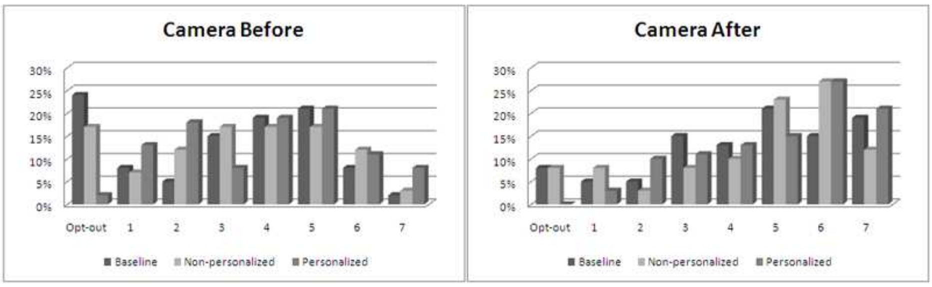


Figure 6.9: Camera ratings before and after, per condition

$p < 0.01$). Post-hoc tests showed that effectiveness was significantly best in the non-personalized condition (Mann-Whitney tests, $p < 0.05$), but comparable between the baseline and personalized condition. Figure 6.8 shows that almost 45% of explanations in the non-personalized condition lead to perfect effectiveness (i.e. Rating1 - Rating2 = 0). In other words H1 is again unsupported, personalized explanations are not most effective.

As such, the correlation between before and after ratings for cameras should be highest in the non-personalized condition as well. Table 6.13 shows that this is indeed the case. Also, it is significant in the personalized and non-personalized conditions, but not for the baseline [11]. It is also worth noting that the correlation found for the non-personalized condition in our experiments in the movie domain (see Tables 6.3 and 6.7) is slightly higher. The correlation in the personalized condition is not highest, also contradicting H1.

We also hypothesized that participants would be more likely to overestimate their ratings of cameras in the baseline condition. Firstly, in Table 6.11 we see that a large number of participants in this condition have opted out. In Table 6.12, we see however that with the opt-out ratings omitted, the initial ratings for cameras are comparable between the three conditions. Figure 6.9 also shows the distribution of these initial camera ratings per condition. That is, there is no significant difference between the explanations in the three conditions w.r.t. persuasion. Moreover, the signed value of effectiveness in Table 6.12 suggests a marginally greater underestimation in the baseline condition. These findings

[11] The same result is found if Spearman's correlation coefficient is used.

Table 6.13: Pearson's correlations between the two camera ratings.

Condition	**Correlation**	**p**
Baseline	0.06	0.70
Non-personalized	0.58	**0.00**
Personalized	0.36	**0.00**

are contrary to our third hypothesis, H3: users are not more likely to overestimate their valuation of items in the baseline condition.

It is surprising that the baseline has reasonably good effectiveness, even if the correlation between before and after ratings for the cameras is not significant. The distribution of initial ratings in the baseline condition (Figure 6.9) suggests that users are less susceptible to persuasion than one might initially think. We return to this when discussing users' qualitative comments.

Are users more satisfied with personalized explanations? (H2)

Table 6.14: Means of the two explanation ratings (opt-outs omitted) in the three conditions.

Condition	**Explanation Before**
Baseline	2.83 (1.44)
Non-personalized	2.38 (1.64)
Personalized	3.27 (1.27)

We compared the users' ratings for the initial explanations (Explanation Before), and found a significant difference between conditions (Kruskal-Wallis, $p < 0.01$). Post-hoc tests support H2, participants were significantly more satisfied with personalized explanations than non-personalized (Mann-Whitney, $p < 0.01$). There was, however, no significant difference between the baseline and the personalized condition.

The second explanation rating (Explanation After) was not analyzed for similar reasons as the experiments in the movie domain: participants seemed to confuse our system with Amazon's, and often rated the reviews on Amazon rather than our explanations.

Limitations and qualitative comments

As we mentioned in our description of the three conditions, the bar chart we used is not comparable to the bar chart used in Herlocker et al. (2000). This is also noted in user comments; *"...doesn't give you information about what kind of customers rated it (a complete newbie wanting to buy a 'point-and-shoot' will rate things differently than amateur buying a SLR)"*;*"No clear indication of the audience that's declared it 'good'. "*

Participants reacted to the number of ratings and the balance between positive and negative ratings in the baseline explanations, which may explain why the baseline performed surprisingly well yet again (w.r.t. effectiveness). Participants were not easily persuaded, and did not rate cameras in this condition persistently highly. For example, a bar chart using too few ratings was considered insufficient information: *"...since it only has the review from six people, it's hard to base my decision on just this explanation...."*;*"Too small all test group for a clear set of results"*.

The majority of reviews were positive even for cameras with many reviews, which by some participants was perceived as poor information as well: *"There are no other opinions except for the people's who are in favor of the camera. This is a poor display of statistics"*;*"everybody cannot possibly rate this good, there have to be some opposers."*. Explanations were taken more seriously when the distribution of ratings was more even; *"The ratings have a larger review base, with some dissenting into "ok", broader review"*.

6.6.8 Summary: Cameras

We found that participants made better decisions in the non-personalized condition, but preferred the personalized explanations. This result is similar to what we found in the movie domain, although we did not expect to find this in an objective, high investment domain. Yet again, the explanations in our baseline conditions fared surprisingly well in terms of effectiveness. The bar chart we used in the baseline did not cause users to overestimate their valuation of items. Rather, this led to more opt-outs as well as many ratings around the mid-point.

6.7 Chapter summary

We found the same result for both the movie and camera domains: participants made better decisions in the non-personalized condition, but preferred the personalized explanations. The replication of results across domains decreases the support for the argument that features for movies are particularly complex and subjective and thus harder to personalize in way that helps users to make decisions. Rather, it gives more fuel to the argument that users do not always understand the factors they use when making decisions. A similar result was found in Ahn et al. (2007) where allowing users to edit their user model ended up decreasing accuracy of recommendations, but increased user satisfaction.

Given the replication of results across domains, we also inquire whether they might be an artifact of our design, i.e. using Amazon as an approximation. If reading Amazon reviews makes the final ratings more positive, an explanation which also leads to overestimation is likely to result in better effectiveness. One possible reason the non-personalized explanations in the movie domain performed better is that they led to a slight overestimation compared to the other conditions. However, we investigated possible causes for overestimation (such as which features were mentioned) and did not find any reason to believe that this was the case (Section 6.4.3). In addition, for cameras, an overestimation of the initial rating would be less likely than for movies - even if the initial valuations of cameras are high, the explanations have no polarity (compared to the movie genres that could belong to preferred, disliked or neutral genres) and should not cause overestimation. For this reason, we believe that the approximation is not the main cause of the repeated results. Nevertheless, in the next chapter (Chapter 7) we describe an evaluation where users got to genuinely try the items.

Chapter 7

Final personalization evaluation

7.1 Introduction

This chapter follows on the results found in Chapter 6, which studied explanation effectiveness in two domains: movies and cameras. These experiments were limited by the approximation we used for evaluating the items: reading online reviews. In particular, the reviews used were positively biased, and so this raises the question if the same results would be found if users actually tried the items. It is possible that non-personalized explanations cause an overestimation that correlates well with the positive bias of the reviews, but that these explanations are not in actual fact effective. In this experiment, participants experience the items.

We chose to conduct this experiment in the movie domain as it is easier, and less costly, to collect suitable materials. This also simplifies the users' valuation of items, as it is easier to let users watch movies than evaluate digital cameras. Given the strong result for baseline explanations in the previous experiments in the movie domain, we also wanted to know how much the title influenced user's ratings of a movie and so included an initial rating (Movie0). The revised procedure is thus:

1. The user rates the item on the basis of the title only **(Movie0)**
2. The user rates the item on the basis of the explanation **(Movie1)**
3. The user tries the item
4. The user re-rates the item **(Movie2)**

A few other details have been changed as a consequence of participants actually trying the items. Primarily the changes concern the choice of materials, which in turn has affected the type of explanations the testbed system generates. See also Section 7.1.2 for more about the selection of movies, and Appendix C for implementational details.

7.1.1 Design

We made as few changes as possible to the design of this experiment compared to the previous experiments in the movie domain. First, we asked participants for information about their preferences. As previously, we restricted the questions to features that can be extracted via Amazon webservices: actors, director, MPAA rating (suitable for children, adult content etc), genre and average rating by other users. We also asked them to select their favorite actors and directors.

Next, participants saw the movie title and rated how likely they were to watch it (Movie0). If they had seen or heard of it before, they were asked to select another movie by clicking a button saying "I might know this movie, please skip to another one". A similar button "I don't want to watch this movie" was also included so that participants would not be forced to watch a movie they did not want to see.

We then gave them an explanation, and asked them to re-rate the movie (Movie1) as well as the explanation (Explanation1). The participants watched the movie, after which they rated the movie for the third time (Movie2), and how much they like the explanation (Explanation2). For ethical reasons, participants had full control over the movies (e.g. they could press the stop and pause buttons), and could terminate the experiment at any time.

Participants were asked to watch as many short movies as was feasible within the duration of an hour, but no more than 3. In a between subjects design, participants were assigned to one of three degrees of personalization, which each reflect the type of explanations they were given.

These were the same as in the the second movie experiment, only the baseline was modified to adjust for the fact that we were using short movies (see Section 7.1.2): e.g. *"This movie is (not) in the top 50 short movies on IMDB (the Internet Movie Database).“*

7.1.2 Materials

As mentioned in the introduction, we chose to use movies rather than cameras. Short movies were chosen over full length features to decrease the experimental duration.

This in turn posed new considerations, such as whether short movies would have sufficiently interesting features such as known actors and directors. Likewise, many short movies do not have certification ratings (such as G, PG and R) which is another frequently used feature.

There were also ethical considerations when selecting materials: for example, participants should not be exposed to sensitive material such as (extreme) violence or sex. Also, to avoid negative experiences participants should not be exposed to material which they really do not want to see, such as movies in strongly disliked genres. For all these reasons, the selected short movies were selected with care.

All the selected movies have some international certification rating, and some have actors (e.g. Rowan Atkinson) or directors (e.g. Tim Burton) that are likely to be known. To aid the selection we used the repository of DVDs at the university as well as the public library in Aberdeen, and the highest rated short movies in the Internet Movie Database (IMDB). 15 movies were selected, out of which 11 are in the top 50 short movies in IMDB. The durations of the movies vary from 3-30 minutes (mean=11.73, StD=10.36).

Table 7.1 summarizes the selected movies. The genres are varied, but the majority of movies belong to the genres comedy (13 movies), animation (11) and children (9). 7 of the movies belong to all three genres. Not all of the movies have a rating, but they have all been individually screened to ensure that they are suitable for 15 years and over (PG-15). Most movies (12) are even suitable for a general audience (G). The movies also vary w.r.t. to other factors, for example some are in foreign languages (English subtitles), the animations have different styles, and three of the movies are black and white.

The small subset of genres is due to our ethical considerations, and arguably this could weaken the experiment. Firstly, if participants dislike the genres comedy, animation, and children, the number of possible movies is greatly reduced. This is highly improbable given that these movies were chosen to minimize offense. Nevertheless, we set as a precondition for signing up that participants would not be averse to these genres.

Secondly, by choosing non-offensive movies, there was a distinct risk that users' ratings of movies would fall close to the middle of the scale and would not be as well distributed as they could be. We argue that for an interesting analysis, it is sufficient that the distribution differs sufficiently from the mid-point, even if it does not make full use of the scale. To confirm this, we conducted two pilot sessions before releasing the experiment to a wider public.

7.1.3 Hypotheses

Our hypotheses H1 and H2 are as before for movies, while H3 regards the camera experiment only. We also have an additional hypothesis due to the addition of an initial movie rating. We believe that the title is not enough information for a user to form an opinion, while the additional information supplied in explanations give added benefit w.r.t. to making a decision. Our hypotheses are therefore that:

Table 7.1: List of short movies

Title	**Genres**	**Duration (minutes)**
Wrong Trousers	Animation, Children, Comedy, Crime	29
Close Shave	Animation, Children, Comedy, Crime	30
Grand Day Out	Animation, Children, Comedy, Crime	23
Feed the Kitty	Animation, Children, Comedy	7
Vincent	Animation, Children, Fantasy	6
For the Birds	Animation, Children, Comedy	3
Mickey's Trailer	Animation, Adventure, Children, Comedy	8
The Rocks	Animation, Comedy, Fantasy	8
Mr. Bean's Christmas	Comedy	4
Rabbit Seasoning	Animation, Children, Comedy	7
Hedgehog in the Fog	Animation, Children, Drama, Fantasy, Mystery	10
Kiwi	Animation, Action, Adventure, Children, Comedy, Drama, Thriller	3
Strange to Meet You	Comedy, Drama	6
Jack Shows Meg His Tesla Coil	Comedy, Drama	7
Somewhere in California	Comedy, Drama	11

- **H1:** Personalized feature based explanations will be more effective than non-personalized and baseline explanations.
- **H2:** Users will be more satisfied with personalized explanations compared to non-personalized and baseline explanations.
- **H4:** Users will able to form an opinion more often after the first explanation (Movie1) than after just seeing the title (Movie0).

7.2 Results

7.2.1 Participants

Participants were recruited from staff and students at the University of Aberdeen, with the selection criteria that they were not averse to the genres children, comedy and animation. 51 participants took part, but 3 were omitted as they specified in their preferences that they did not want to be recommended one of the above genres (e.g. comedy), and were therefore prematurely directed to the final debriefing screen. The remaining 48 were equally distributed between the three conditions (16,16,16). The mean age of participants was 26.17 (StD=7.24), and were roughly equally distributed w.r.t. gender (21 male, 27 female).

7.2.2 How do titles compare to explanations? (H4)

Table 7.2: Opt-outs for movie ratings (%)

Condition	**Movie0**	**Movie1**	**Movie2**
Baseline	51.1%	28.9%	0%
Non-pers.	20.9%	7.0%	0%
Personalized	34.1%	11.4%	0%
Average	35.6%	15.9%	0%

Opt-outs

Looking at the percentages of opt-outs in Table 7.2 we see that, on average, participants opted out 35.6% of the time for Movie0, compared to 15.9% after receiving an explanation (Movie1). In Figure 7.1 we see the change in opt-outs for the three movie ratings across the conditions, and note the noteworthy decrease of opt-outs from Movie0 to Movie1 in all three conditions. This suggests that explanations do help users to make decisions. We also investigated whether the difference in opt-outs between Movie0 and Movie1 is significant. In order to do this, we recoded the movie ratings into two binary values: “0” for opt-outs, and “1” when a rating was given. We then compared the distribution of 0’s and 1’s between Movie0 and Movie1, in each condition. The difference in a binomial sign test was significant at $p < 0.05$ for all three conditions. Surveying the direction of changes, H4 is confirmed - more participants were able to make decisions with the explantions than with just the title. In Section 7.3 we also discuss if the large number of opt-outs for Movie0 in the *baseline* could be an artifact of the movies shown to the participants, individual differences, or something else.

Mid-scale ratings

Table 7.3: Means of the three movie ratings (excluding opt-outs)

Condition	**Movie0**	**Movie1**	**Movie2**
Baseline	4.36 (0.95)	4.28 (0.81)	4.76 (1.67)
Non-personalized	4.12 (1.67)	4.45 (1.53)	4.58 (1.88)
Personalized	3.86 (1.23)	4.31 (1.26)	4.93 (1.86)

It is also worth surveying the proportion of ratings in the middle of the scale (value=4). While these are not as strong an indicator of (lack of) informativeness as opt-outs, a larger proportion in the middle of the scale could suggest that participants had difficulty in forming a strong opinion.

Table 7.3 summarizes the mean ratings, and in Figure 7.2 we see the distribution of ratings for the three conditions for all three movie ratings. Here we see that Movie1 and Movie2 are distributed beyond the mean rating of 4, suggesting that participants are able to form opinions across the scale. Movie1 and Movie2 ratings are also skewed toward the higher end of the scale; we can see that there are more ratings of value 4 or above. This skew is not completely surprising given that we had selected movies that were unlikely to cause offense, as well as avoided genres and movies that participants did not want to

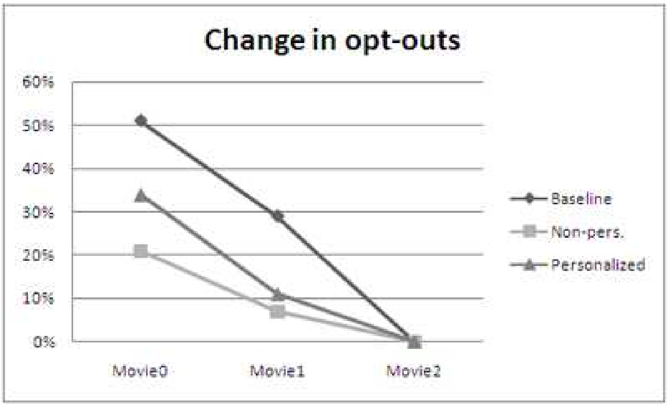
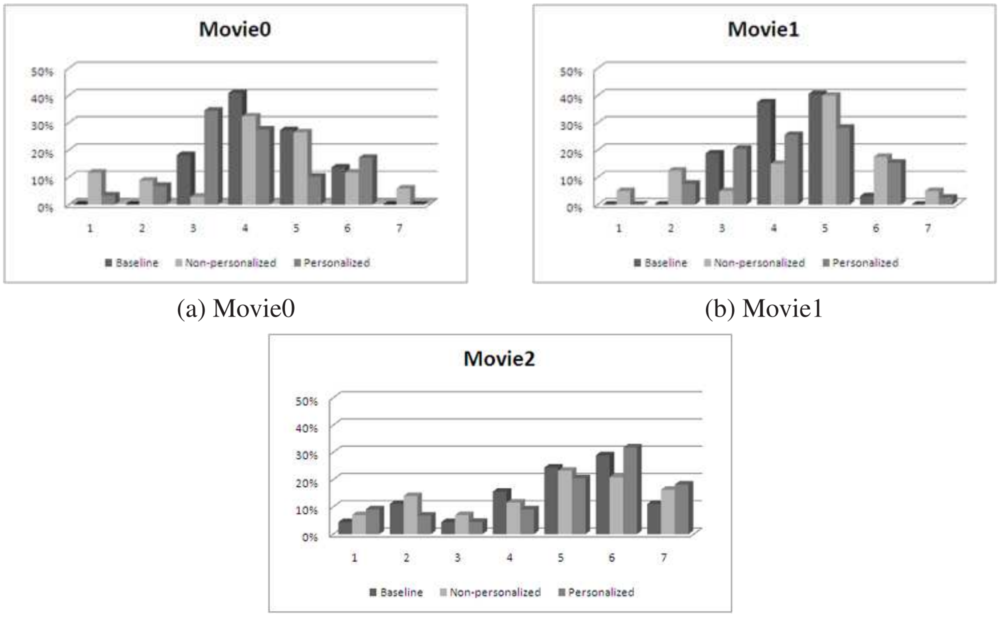


Figure 7.1: Change in opt-outs between the three opt-out ratings, including opt-outs

see. For Movie0 however, there is a relatively large proportion of opt-outs. There are also more mid-scale ratings of 4 in the non-personalized and personalized conditions for Movie0 than for Movie1 and Movie2. The large number of opt-outs and mid-scale ratings suggest that users struggle to specify an opinion with the title alone.

We have seen that there are slightly less 4s for Movie1 in the non-personalized condition (not significant). In Section 7.3 we discuss if this finding could be an artifact of the movies shown to the participants, individual differences or something else.



(a) Movie0

(b) Movie1

(c) Movie2

Figure 7.2: Distribution of movie ratings, the distribution is considered with regard to the *percentage* of ratings in each condition

7.2.3 Are personalized explanations more effective? (H1)

Table 7.4: Mean effectiveness with "opt-outs" omitted, and Spearman's correlations (*1-tailed*) between the two movie ratings.

Condition	**Effectiveness (absolute)**	**Effectiveness (signed)**	**Correlation**	**p**
Baseline	1.09 (1.00)	-0.41 (1.43)	0.44	**0.01**
Non-pers.	1.78 (1.37)	-0.08 (2.26)	0.27	**0.05**
Personalized	1.69 (1.08)	-0.41 (1.98)	0.24	0.07

We see the mean effectiveness in each condition summarized in Table 7.4. Looking at the signed effectiveness we see that in all conditions the explanations led to a slight underestimation. Surprisingly, explanations in the baseline condition lead to the "best" effectiveness ($p < 0.05$, Kruskal-Wallis) according to this correlation measure. This

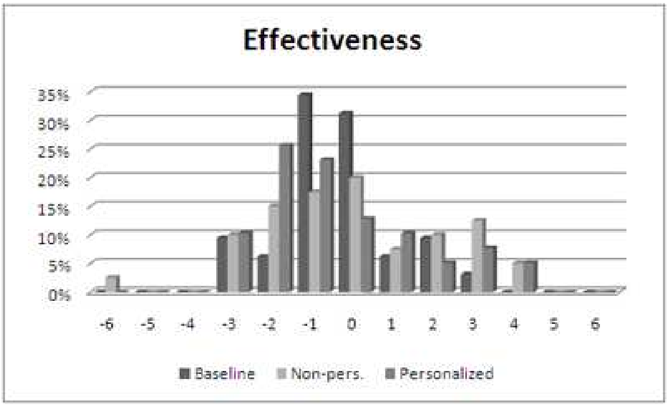


Figure 7.3: Distribution of effectiveness, excluding opt-outs

finding is in stark contrast to the large number of opt-outs in this condition, which indicate that baseline explanations are clearly not helpful more than half of the time.

One possible explanation for the strong correlation is that the baseline explanations biased the participants toward high ratings, as most of the short movies were in the top 50 in IMDB. As the selection of movies was guided by being acceptable to users, this also was likely to lead to a large proportion of high ratings. In Figure 7.2 (Movie1) we see a skew toward high ratings in all conditions, but this skew is not the most severe in the baseline. Movie1 ratings are comparable for all conditions (no significant difference for Kruskal-Wallis), suggesting that the baseline is not more strongly skewed toward positive ratings. We also surveyed how many of the shown movies in the baseline were in the top 50, and found that the relative proportion was comparable (44.4% were in the top 50, and 55.6% were not). So, it seems that the large number of mid-range ratings is a more plausible explanation than a skew toward positive ratings. (We explain why a rating distribution with many mid-range ratings may lead to misleading measurements of effectiveness in Section 7.4.)

Table 7.4 summarizes the Spearman's correlation (*1-tailed*) between Movie1 and Movie2 for all three conditions. The correlation is strongest in the baseline, weak in the non-personalized condition, and not significant in the personalized condition.

7.2.4 Are users more satisfied with personalized explanations? (H2)

Table 7.5: Explanation ratings, means and opt-outs.

	Explanation1		**Explanation2**	
Condition	**Mean (StD)**	**Opt-outs**	**Mean (StD)**	**Opt-outs**
Baseline	2.55 (1.43)	14.6%	2.89 (1.60)	2.1%
Non-pers.	3.51 (1.61)	4.5%	3.53 (2.00)	0.0%
Personalized	3.21 (1.46)	4.3%	3.16 (1.83)	2.2%

We hypothesized that participants would prefer personalized explanations to non-personalized and baseline explanations. First, we look at the opt-out rates. In Table 7.5 we see that while the opt-out rates for the explanations in the two feature based conditions are comparable, the opt-out rate for explanations in the baseline is much higher.

Next, we investigate if participants preferred the personalized explanations over the explanations in the other two conditions. First, we look at the initial explanations ratings (Explanation1). A Kruskal-Wallis test shows a significant difference between the ratings for the baseline and non-personalized explanations ($p < 0.05$), but not for the ratings between the personalized condition and the other two conditions (although the mean rating for Explanation1 indicates that participants preferred non-personalized explanations over personalized ones). This contradicts our previous findings, where personalized explanations were preferred in both the camera and movie domain. We will discuss this further in Section 7.3.2.

In our previous experiments we were not able to compare the ratings for Explanation2 as participants confused our testbed with Amazon itself. As this confusion no longer is a factor in the experiment we can study the participants opinion of the explanations after watching the movie. To our surprise we found no significant difference between the conditions, but note that the non-personalized explanations still have the highest mean rating, followed by the personalized explanations.

7.3 Discussion

A larger proportion of participants opted out for Movie0 in the baseline condition compared to the other two conditions. Participants had at this stage not yet been given explanations. We also see that the mean Movie0 rating is lowest in the personalized condition, despite users only being showed the title. We imagine two possible reasons for this: either the titles differ somehow, or there are notable individual differences between participants. Let us first have a look at the titles.

Difference in shown titles?

If the movie titles shown to participants in the baseline were less informative (revealing less features such as the character in “Mr. Bean’s Christmas”) than in the other two conditions, this could explain the presence of more opt-outs. We investigated the frequency of movie titles in the different conditions. Due to random selection these differed slightly between conditions, but there was no noteworthy difference in the distribution of titles between the three conditions. Roughly the same movies were shown in all three conditions, and though the numbers differed they were largely similar. The selection of movies does not seem to explain the number of opt-outs.

Differences between users?

It is harder to discuss the effects of individual differences as participants only rated up to three movies, but we did survey the rating patterns of individual participants. While it is evident that participants in the baseline on average opted out more for Movie0, it does not follow that all participants opted-out for all the titles they rated: there were many cases where participants opted out for some of the movies, but not all. It is however possible that once a participant opted-out once, they were more likely to opt-out for the remaining movies they saw. Likewise, the lower ratings for Movie0 for the personalized explanations may have "carried over" multiple movies. However, while individual difference might explain the large number of opt-outs for Movie0 in the baseline, for Movie1 we believe that other factors are at play as well.

While the opt-out rate for Explanation1 could be explained by participants in this condition being more likely to opt-out on any question, the low ratings for Explanation1 (see also Section 7.2.4) would support the hypothesis that baseline explanations are considered less informative. This is corroborated by the large proportion of opt-outs in the baseline for our previous experiments. The general large number of opt-outs in the baseline is likely to have been affected by individual differences, but for Movie1, is likely to be due to lack of informativeness (of the baseline) as well.

7.3.1 Top 50?

For those participants that did not opt out, the baseline explanations performed the best w.r.t. effectiveness. We are therefore curious to see whether or not a movie is in the top 50 on IMDB (in short denoted as Top50) affects Movie1, and effectiveness in this condition. The Pearson Chi-square correlation between Movie1 and Top50 is not significant although there is a weak trend in this direction. There is also no correlation between Top50 and effectiveness. That is, information about popularity in this case does not seem to influence effectiveness.

7.3.2 Qualitative comments

The correlation between Movie0 and Movie1 suggests that participants may use the movie title to inform their opinion. Indeed, this is reflected in the participants' comments: *"..I considered what it might be like based on the title and the genre (personalized condition)","I thought the title (Rabbit Seasoning) suggested rabbits being killed but the light hearted nature of the film made me enjoy it more than I thought I would."*

Secondly, we were surprised to see that personalized explanations in this experiment were not rated significantly higher than non-personalized explanations. A large proportion of the explanations in the personalized condition describe the actors, but the information is not as useful to the participants as they might have anticipated. The names used were largely unknown to participants, as it was harder to find short movies with known actors. For animations in particular, participants saw the names of actors whose voices were recorded, rather than the character they played (e.g. Bugs Bunny or Gromit). This could have decreased the satisfaction with the personalized explanations as participants were less likely to recognize them: *''Also without knowing who the star is, this could still not mean a lot to the description.''*.

Likewise, some participants complained they did not recognize the director: *"the director's name is even less recognizable than the actors' names..."*;*"As I don't know the director, the rest of the description could easily belong to a totally different movie."*

We investigated the frequency of favorite actors and directors in the user profiles. While sparse, the frequency is similar to that of previous experiments. That is, people specify a comparably small number of favorite actors and directors in all experiments. There is however a difference in that participants recognized the same few names in this experiment, potentially making it less likely that they come across a movie where it would be relevant to mention this name. E.g. most participants recognized Rowan Atkinson, but few other actors, and he only stars in one of the used short movies. Likewise, if the actors and directors mentioned were somehow misleading (e.g. if their favorite actor plays a weak role), this would also impede effectiveness.

Participants also commented on factors that were not considered so important in our focus groups, but which may have been identifying for the movies they were shown. For example, while our focus groups participants said they did not care about movie studio, this does affect the style of animation: *"...pretty much what I'd expect from a Pixar movie."*;*"Unlike the last movie I was not expecting a Walt Disney film..."*.

7.4 Summary

The results of this experiment once again highlight the importance of selecting relevant evaluation criteria. While the baseline explanations were found to be the most effective, they also had the lowest satisfaction, and led to most opt-outs and ratings in the middle of the scale.

We found that the title alone could lead to the same rating as a simple explanation, however it also often leads to opt-outs. The difference between the Movie0 and Movie1 ratings offer an argument in favor of explanations in recommendations: participants in all three conditions opted out half as much after receiving an explanation.

Both feature based explanations were initially (Explanation1) rated higher than baseline explanations, but only the difference between non-personalized and baseline explanations was significant. We believe that the weaker result for personalized explanations in this experiment compared to our previous experiments is due to the restricted choice of materials. It was difficult to find short movies with known actors and directors, and despite a conscious effort to use movies with known features, this type of overlap is likely to have been sparse. That is, it was less likely that participants would encounter familiar actors and directors in this experiment compared to our previous experiments in the movie domain.

This experiment also brings into light two situations where our evaluation metric for effectiveness could fail:

1. **If a large proportion of ratings fall on the middle of the scale.**
   Firstly, mid-scale ratings are ambiguous, we do not know if users are selecting this option because they cannot make a decision, or because they feel that the item is precisely “ok” (i.e. neither good, nor bad). The presence of an opt-out option helps clarify this, but only partially, as participants may supply a neutral value when they do not have enough information to form a polarized opinion. For example, for the baseline Movie0, and Movie1 ratings we saw that participants gave more midscale (“4”) ratings than in other conditions: this was more likely due to a lack of information than a large proportion of movies that were precisely ok.
   Secondly, explanations that cause a large proportion of mid-scale ratings (for Movie1) are likely to lead to better effectiveness than explanations that result in more equally distributed ratings. Even the most extreme changes in opinion can only be as big as half of the scale. A wider distribution of the initial ratings, assuming that the after ratings are normally distributed around the middle of the scale, is likely to lead to greater divergences. Thus, smaller errors might then (at least in part) be due to the poor distribution of the before ratings, rather than better effectiveness.
   It is also arguable that if the initial ratings (Movie1) are random but also follow a normal distribution (our current assumption for Movie2), the *signed* average difference would be 0. This is why we highlight the importance of measuring the *absolute* value of the difference.

2. **If the explanations are biased in the same direction as the data.**
   Following on the previous point, we can imagine an extreme scenario where a recommender system gives explanations that result in a large proportion of neutral ratings, and only recommends “safe” items that usually score around the middle of

the scale. These explanations might appear to be effective[1], but are not likely to be particularly informative. This does not mean that the explanations are generally effective, as they would be misleading for non-neutral recommendations.
In addition, in our experiments we have seen that neither the before or after ratings were centered around the middle of the scale. In this case, it makes more sense to consider the mean rating (e.g. 5/7) for the after distribution (Movie2) rather than the middle of the scale (e.g. 4/7). That is, false effectiveness may be found if there are many initial ratings (Movie1) around the value that is the mean of the after ratings (Movie2). We can imagine explanations that inflate the initial valuation of items and only recommend the most popular items; or explanations that devalue items and only recommend unpopular item. In these cases our metric for effectiveness may result in high correlations, and a mean difference of 0 between the before and after ratings. However, this does not mean that the explanations are effective. For this reason, the underlying distribution of ratings should be presented alongside any measurement of effectiveness.

We caution that neither of these situations per default imply a failed metric. The items may in fact be just ok, and a system that helps to identify this correctly should not be classified as faulty. Likewise, "biased" explanations may be suitable if this fits the data e.g. positive explanations for items that the user is predicted to like. Baseline explanations like ours may make sense if they are based on many previous user opinions as is the case with the Internet Movie Database (IMDB). However, it would be prudent to assume that explanations cannot be ported between datasets, or domains without careful consideration. Any study using the same metrics for effectiveness would need to study the underlying distribution as well. For these reason we would encourage replication of this experiment with other materials and in different domains, to confirm which of our findings carry beyond our small selection of materials.

The limited dataset, and sparsity of known features mentioned above may have contributed to lower effectiveness and satisfaction for the personalized condition, but would then be a consideration for any dataset where this type of sparsity is likely: where the features are likely to be misleading or uninformative. Again we found that non-personalized feature based explanations were more effective than personalized, suggesting that users might not always be the best judges of which information would be most useful to them in these scenarios.

Last, but certainly not least, we consider the effect of letting users watch the movies contra reading movie reviews on Amazon. In our case it is difficult to separate the effects of material choice from the effects of the change in design. The baseline explanations

[1] Assuming a normal distribution of the after ratings, the difference for another fixed but random value (e.g. if all Movie1 ratings were equal to 5) would be larger than using the middle of the scale.

were the most effective in this experiment, but this does not seem to be due to an initial overestimation because Movie1 ratings were comparable between conditions (see Section 7.2.3). It is more likely that the popularity of short movies is a better predictor than for long movies. Indications of this can be seen in Section 7.3.1.

Our previous experiments in two domains used Amazon as a data source and led to repeated results. One could therefore also argue that using Amazon as a dataset leveraged the results for feature-based explanations w.r.t. effectiveness in our previous experiments. That is, reading reviews on Amazon caused an overestimation that correlated well with the (also overestimated) valuation of items after reading explanations. Although not due to bias as in the Amazon reviews, the average Movie2 ratings in this experiment are however also high. This bias would therefore also benefit from a presumed positive skew caused by feature-based explanations, but in this experiment no such bias is evident as the feature-based explanations show worse effectiveness. Our suggestion is therefore that the dataset was more likely to have affected our results than the change of design, in particular with regard to satisfaction. It is also worth considering that the baseline just was more effective, because it is a good data source (given the large number of user ratings available on IMDB). Another alternative explanation for our results is that while the baseline was not the best possible explanation (inferring from the large number of opt-outs), the type of personalization we used in the personalized condition does not contribute to effectiveness. Naturally, further similar experiments with alternative datasets would be required to confirm that this really is the case.

Chapter 8

Conclusions and future work

In this chapter we summarize our findings, and the answers to our main research questions mentioned in the introduction, Section 1.1. First, we address the question of why we should explain, or whether there is any point in explaining recommendations at all (Section 8.1). Then, we discuss if the type of explanations that we offer to users matter, or rather if our personalization of explanations increased their effectiveness (Section 8.2.1). Given that our results may have been due to the choice of methodology, a large portion of this section is dedicated to methodological discussion. Next we discuss how to best measure the effectiveness of explanations. In Section 8.3, we summarize the lessons we have learned about the used metric for effectiveness, and its relation to the underlying data. Finally, we conclude with suggestions for future work in Section 8.4.

8.1 Should we explain at all?

For a recommender system aiming at user satisfaction rather than decision support, well formed explanations can contribute positively: we saw that personalization of explanations does increase satisfaction compared to a baseline, although we have not compared a system with explanations with one without. Table 8.1 summarizes participant's satisfaction with the explanation, by condition, for each of the four related experiments: MoviesI, MoviesII, Cameras and Final Eval. The experiments *MoviesI* and *MoviesII* are described in Sections 6.4 and 6.5, and describe our initial investigations in the movie domain, controlling for some possible confounding factors in MoviesII. *Cameras* repeats the experiment in a second domain, and is described in Section 6.6. The final evaluation (*Final Eval*) uses a different methodology whereby participants tried the items rather than just approximating their valuation as was done in the first 3 experiments. This experiment is described seperately in Chapter 7.

Table 8.1: Initial satisfaction with explanations (on a scale from 1-7, 1=really bad explanations, 7=really good explanations)

Condition	**MoviesI**	**MoviesII**	**Cameras**	**Final Eval** (1-tailed)
Baseline	2.38 (1.54)	-	2.83 (1.44)	2.55 (1.43)
Non-pers.	2.50 (1.62)	2.72 (1.68)	2.38 (1.64)	3.51 (1.61)
Pers.	3.09 (1.70)	3.31 (1.55)	3.27 (1.27)	3.21 (1.46)

Effectiveness was comparably strong for all of our explanations, in all experiments: the correlations for before and after ratings were significant for feature-based explanations, and the mean absolute effectiveness was reasonable for all explanations. Table 8.2 summarizes the change of opinion, where we hope to minimize the change of opinion, for our four experiments. We see that average change is on the magnitude of 1 scale point on a 7 point scale. Effectiveness was generally slightly better for movies than for cameras. Table 8.2 summarizes the correlation between before and after item ratings for all four experiments.

This suggests that explanations, our baselines included, can offer relevant (albeit limited and imperfect) information, with the caveat that baseline explanations have led to more opt-outs for the initial rating in all four experiments. See also Table 8.4 for a summary of opt-outs[1]. Part of the strong result for baseline explanations (when participants did not opt out) may have been due to the presence of a title for movies, but the replicated

Table 8.2: Mean (absolute) effectiveness with "opt-outs" omitted, per experiment.

Condition	**MoviesI**	**MoviesII**	**Cameras**	**Final Eval**
Baseline	1.38 (1.20)	-	1.77 (1.50)	1.09 (1.00)
Non-pers.	1.14 (1.30)	0.96 (0.81)	1.14 (1.32)	1.78 (1.37)
Pers.	1.40 (1.20)	1.33 (1.27)	1.88 (1.34)	1.69 (1.08)

Table 8.3: Pearson's correlations ($p < 0.01$, 2-tailed unless otherwise specified) between before and after ratings of items, with "opt-outs" omitted, per experiment.

Condition	**MoviesI**	**MoviesII**	**Cameras**	**Final Eval** (1-tailed)
Baseline	0.43	-	0.06 ($p = 0.70$)	0.44 ($p < 0.01$)
Non-pers.	0.65	0.79	0.58	0.27 ($p < 0.05$)
Pers.	0.58	0.56	0.36	0.24 ($p < 0.07$)

[1] The lack of data for the baseline condition for MoviesII in Table 8.3 reflects a large opt-out rate,and extremely short duration times. See also Section 6.5.

finding for cameras is promising (see also Table 8.2): explanations (with or without titles) can help in making decisions.

In one of the experiments (described in Chapter 7), we allowed participants to rate the title alone, and then rate the item again once they saw the explanation. The number of opt-outs decreased *significantly* once participants received an explanation. That is, explanations also add to effectiveness in terms of increasing the number of items that users feel that they can evaluate.

Table 8.4: Percentage of opt-outs, Item Before, per experiment

Condition	**MoviesI**	**MoviesII**	**Cameras**	**Final Eval**
Baseline	8.8	55.6	23.9	28.9
Non-pers.	7.2	4.3	16.7	7.0
Pers.	3.1	15.2	1.6	11.4

8.2 Personalization

8.2.1 Personalization - summary

We ran three initial experiments using our testbed explanation generation system in two domains, and found that our method of personalization hindered effectiveness, but increased satisfaction with explanations. However, these experiments were based on an approximation of effectiveness where participants read review for items rather than trying them. In our final evaluation, participants were able to watch the movies. In this case, the opt-out rate for the baseline explanation (Movie1) was much higher than for the other two conditions (see also Table 8.4). For the remaining movie ratings, both feature-based explanation types were *less* effective than baseline explanations. The non-personalized explanations were most preferred by participants. In the final evaluation, we believe that the personalized explanations were disadvantaged by the change to short movies, and the nature of the baseline (in the Internet Movie Database ratings for short movies may be more informative than for long movies). While the results for the approximated and true effectiveness experiments are not entirely consistent there are three conclusions that can be drawn for all experiments:

1. Contrary to our initial hypothesis, personalization was in most cases clearly *detrimental* to effectiveness.

2. Users are more likely to be more satisfied with feature-based than baseline explanations. If the personalization is perceived as relevant to them, then personalized feature-based explanations are preferred over non-personalized.

3. User satisfaction is also reflected in the proportion of opt-outs, which is highest for the baseline explanations in all experiments. This was the case despite the different types of baselines used in the two domains.

Our findings are similar to those of Reiter et al. (2003), who did not find that personalized smoking cessation letters helped smokers more than non-personalized letters. We suggest four possible reasons why our personalization did not increase the effectiveness of explanations:

1. Personalization is not effective in a particular domain.
2. Personalization might have been effective if we had used the right type of explanations e.g. using a deeper user model, longer explanations etc.
3. The system does in fact generate effective personalized texts, but the experiments failed to detect this because of design issues.
4. Users cannot always correctly judge what information they they need in order to make decisions.

All of these explanations are relevant to our experiments, and we will address each point below.

8.2.2 Domain

In our experiments, personalized explanations performed worse w.r.t to effectiveness in two domains. The two domains were chosen because they were each others polars on the two dimensions: high vs. low price, and subjective vs. objective valuation. Therefore, we believe it is unlikely that the initial domain choice of movies (e.g. it is harder to evaluate movies objectively) is the reason that non-personalized explanations were found to be more effective than personalized. It may of course be the case that our form of personalization is not effective for either domain, but would be for a third. However, given that personalized explanations were less effective in both, it is more likely that the flaw lies elsewhere.

8.2.3 "Wrong" explanations and models

We used a very simple user model, so we should certainly not infer from our results that *any* form of personalization is damaging for decision support. If we had made more use of the information about the user and/or item, we might have been able to generate more effective personalized explanations. Likewise, perhaps explanations cannot be really effective unless they are part of a process of interaction where the user learns about existing and competing options, as is the case in conversational recommender systems (Felfernig et al., 2008; McSherry, 2005; Reilly et al., 2004b). That is, a user might change their mind once they learn about competing options, and consequently refine their explicit requirements to better reflect their genuine preferences (which they previously might not have been able to identify or formulate). We are particularly interested in whether we used the right information, and if length was an influential factor.

Right information

We know from our user studies in Chapter 5 that users would have preferred different types of information than the ones we supplied in our explanations. It is possible that using simple features compromised the effectiveness of the explanations. In our experiments however, there was a trade-off between what was readily available on Amazon, and the kind of information users may have preferred to make decisions.

We mentioned in Section 6.3 that our aim was to generate explanations that could realistically be integrated into an existing commercial system. Following on this, we highlight the choices we made for the personalization in particular, as well as which limitations these choices have imposed on the generated explanations.

We were limited by both the number and type of features that were available from the Amazon Webservices (and similar limitations are likely to occur with other existing commercial services that span a large number of domains). Some features, such as what kind of mood a movie is suitable for, or the complexity of the script, cannot easily be extracted. This limited which and how many features we could use.

Other features were not as complete as they could be. For example, we knew from focus groups that users place more importance on if an actor or actress plays a particularly noteworthy, or famous, role (see also Chapter 5). In our explanations we mention leading actors according to user preference, but it is by far more difficult to describe this feature in a deeper manner.

While natural language processing techniques can be adapted for both these (and similar) cases, this would require a deviation from the main focus of this thesis. The point was whether personalization in a realistic scenario, using existing metadata and APIs, could aid decision support, and not to learn how to best e.g. extract in-depth information about actors from reviews. Likewise, even if we assume that complex features such as those mentioned above could be deduced about movies in an off-the-shelf manner, there remains the problem of (explicitly) inferring the equivalent user interest, e.g. users who only like Al Pacino when he plays well, and only in action movies, from user ratings alone. In contrast, content-based algorithms, or hybrid algorithms that are partly content-based (e.g. Symeonidis et al., 2008), which consider simple features such as actor and director names already exist. Likewise, we found strong and repeated indications that people varied with regard to which features they found important (see Chapter 5), as well as a precedent in the literature for the effect of personalization on persuasion (Chapter 4).

We also considered other types of personalization such as the changing the medium of presentation e.g. text vs. graphics, or whether to describe movies in more or less detail, but decided not to pursue these questions further, see also Sections 4.6 and 5.4.

Length and completeness

There is also a risk of over-personalization for explanations as short as ours. That is, the personalized information is offered at the cost of other relevant information, and longer explanations that considered more item features, or considered them in more detail, may have performed better. In Chapter 5 we saw that longer reviews were preferred to shorter, so length may well affect effectiveness of explanations.

In the first experiment in the movie domain, we saw that participants wanted to know all the genres each movie belongs to. It was not enough for the participants to know that the movie was in a genre they wanted, or did not want, to see. It is fully possible that participants consider more information than can be "formally" justified in the final decision. The user might either not be able (or willing) to formulate their criteria well (see also Section 8.2.5), or just not be aware of the options yet. This later would be particularly relevant in domains where the user first needs to learn more about the domain as observed by Felfernig and Gula (2006): different item features may change importance as the participant observes competing options. However, whether length would affect effectiveness, satisfaction, or both, for explanations (rather than reviews) is still an open question.

8.2.4 Design

With regard to design, there are a number of factors to consider such as our choice of materials, the used approximation (reading online Amazon reviews), and the effect of explicitly asking participants for their preferences.

Material selection vs. design change

In the final evaluation, if we disregard the large number of opt-outs for the movie ratings in the baseline condition, these explanations lead to better values for effectiveness than explanations in the two feature-based conditions. We cannot be sure why this result is different from the results for the other experiments. Two likely candidates are a) the change from an approximation (such as reading online reviews) to a true experience of the items, and b) the effects of material choice.

For the approximations one could argue that the non-personalized explanations somehow lead to a positive bias (one argument could be that there are more mentions of average ratings, that tend to be high, in these explanations) that aligned well with a pre-existing positive bias on Amazon. Surveying the distribution of movie ratings, and the nature of the chosen materials for the final evaluation, it seems as if the materials had a larger effect however. The initial movie ratings (Movie1) do not differ significantly between conditions in any of our experiments, and the final ratings in the final evaluation (Movie2) are positively skewed. The positive skew in Movie2 should have leveraged the non-personalized explanations in the same way as in the previous experiments. Since this is not the case, we can not claim that a positive skew in after ratings overly leveraged the non-personalized condition in our previous experiments.

8.2.5 Explicit vs. implicit preferences

Our user model was also influenced by being based on explicit preferences (see also Section 8.2.5). In the final evaluation, the “personalized information” using actor names was not useful to the participants. In this experiment there were many animated movies, and for these we listed the names of the actor voices rather than character names. It might be the case that users do not understand the factors they use when making decisions, or how they relate to the available options. In these types of cases, users might benefit from the system building the user model based on implicit preferences. A related result was found by Ahn et al. (2007) where allowing users to edit their user model decreased the accuracy of recommendations, but increased user satisfaction.

In terms of satisfaction with explanations, it is valid to argue that participants’ expectations of the explanations were influenced by the initial questionnaire. Perhaps participants gave higher ratings to personalized explanations because they expected them to be personalized. In our experiments, we have not differentiated between satisfaction due to the participant feeling that the explanation helped them to make a decision, and satisfaction due to the fact that they feel that the system is responsive and considering their preferences.

It is worth noting however, that in the final evaluation the explanation ratings after watching the movies were comparable between conditions. It is likely that participants adjusted their ratings of the explanations with regard to how helpful they were, giving non-personalized and baseline explanations higher ratings than before.

8.3 More thoughts on effectiveness

Through our series of experiments with the testbed, we have learned a few things about measuring effectiveness. Firstly, it is worth to differentiate between over- and underestimation as they are perceived differently by users (perceived helpfulness). In Section 4.5 we describe a study in which users considered overestimation to be less helpful than underestimation, and overestimation to be less helpful in high investment domains than in low investment domains. We also found that the same distance between points on a scale cannot be assumed to be equal.

Secondly, the validity of the used metric (see Section 4.3.4) is dependent on the underlying dataset: the distribution of ratings needs to be presented together with the other metrics of effectiveness such as absolute or signed mean. E.g. if participants consistently use the middle of the scale to indicate that they have no real opinion about a movie, this may seem like perfect effectiveness, when in fact it's not (see also Section 7.4 for a discussion). In addition, effectiveness may not capture skews in data, e.g. if the explanations are persuasive and result in high initial ratings, or degrade the items and result in low ratings. For example, if the explanations degrade the items, and the dataset consists of items most users will consider as bad, effectiveness will be good. This does not make these explanations universally helpful.

Finally, this metric does not consider opt-outs. As mentioned previously in this chapter, the baseline explanations in our experiments suffered from a large number of opt-outs even if the effectiveness (measured as the absolute mean of the difference between the before and after ratings) was seemingly comparable with other conditions. For an explanation to be effective, it has to at the very least elicit some sort of rating (preferably one that reflects the user's preferences). An explanation that cannot help elicit *any* rating, by definition leads to poor effectiveness, and moreover is likely to result in user satisfaction so low that the system is likely to lose the user.

8.4 Future work

We found that while personalization often increased satisfaction, contrary to our hypothesis, it was detrimental to effectiveness. It may be the case that personalization in general does not increase effectiveness. In this chapter we considered if this result is more specific to our studies, and discussed how our choice of experimental design, and type of explanations generated may have led to this surprising result.

We encourage further studies with more complex or simply longer explanations (e.g. based son deeper user models) and using a different design (e.g. different materials). Our suggestions for related future work also include using an implicitly learned user model given that users may not always know what information they need to make accurate decisions.

While the independence from a particular recommender system has allowed us to run controlled experiments, it would also be interesting to conduct studies with a live recommender system. That way one could for example conduct longitudinal studies such as the effect of explanations on trust, and see in which situations trust increases and decreases over time.

In addition, other researchers are starting to find that explanations are part of a cyclical process. The explanations affect a user's mental model of the recommender system, and in turn the way they interact with the explanations. In fact this may also impact the recommendation accuracy negatively (Ahn et al., 2007; Cramer et al., 2008b). For example Ahn et al. (2007) saw that recommendation accuracy decreased as users removed keywords from their profile for a news recommender system. Understanding this cycle will likely be one of the future strands of research.

It also remains an open question how much personalization of presentational choices would affect actual (rather than perceived) effectiveness. So, in conclusion, while this is an exhaustive and self-contained piece of work, there are many interesting avenues still left to explore!

Bibliography

Adomavicius, G. and Alexander Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transations on Knowledge and Data Engineering*, 17:734–749.

Adrissono, L., Goy, A., Petrone, G., Segnan, M., and Torasso, P. (2003). INTRIGUE: Personalized recommendation of tourist attractions for desktop and handheld devices. *Applied Artificial Intelligence*, 17:687–714.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. I. (1996). Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining*, pages 307–328.

Ahn, J.-W., Brusilovsky, P., Grady, J., He, D., and Syn, S. Y. (2007). Open user profiles for adaptive news systems: help or harm? In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 11–20, New York, NY, USA. ACM Press.

Ajzen, I. and Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, 84(5):888–918.

Andersen, S. K., Olesen, K. G., and Jensen, F. V. (1990). *HUGIN—a shell for building Bayesian belief universes for expert systems*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Ardissono, L., Gena, C., Torasso, P., Bellifemine, F., Difino, A., and Negro, B. (2004). *Personalized digital television*, chapter 1, pages 3–26. Kluwer Academic Publishers.

Armengol, E. and Plaza, E. (1994). A knowledge level model of case-based reasoning. In *EWCBR '93: Selected papers from the First European Workshop on Topics in Case-Based Reasoning*, pages 53–64, London, UK. Springer-Verlag.

Balabanovic, M. and Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM*, 3:66–72.

Basu, C., Hirsh, H., and Cohen, W. (1998). Recommendation as classification: Using social and content-based information in recommendation. In Rich, C. and Mostow, J., editors, *15th National Conference on Artificial Intelligence (AAAI-98)*, pages 714–720.

Bederson, B., Shneiderman, B., and Wattenberg, M. (2002). Ordered and quantum treemaps: Making effective use of 2d space to display hierarchies. *ACM Transactions on Graphics*, 21(4):833–854.

Bell, A. (2002). *The Language of News Media*. Blackwell Publishers.

Bell, R. M. and Koren, Y. (2007). Lessons from the netflix prize challenge. *SIGKDD Explor. Newsl.*, 9(2):75–79.

Bennett, S. W. and Scott., A. C. (1985). *The Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, chapter 19 - Specialized Explanations for Dosage Selection, pages 363–370. Addison-Wesley Publishing Company.

Bilgic, M. and Mooney, R. J. (2005). Explaining recommendations: Satisfaction vs. promotion. In *Proceedings of the Wokshop Beyond Personalization, in conjunction with the International Conference on Intelligent User Interfaces*, pages 13–18.

Billsus, D. and Pazzani, M. J. (1999). A personal news agent that talks, learns, and explains. In *Proceedings of the Third International Conference on Autonomous Agents*, pages 268–275.

Bleich, H. L. (1972). Computer-based consultation electrolyte and acid-base disorders. *The Americal Journal of Medicin*, 53(3):285–291.

Breese, J. S., Heckerman, D., and Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. pages 43–52.

Bridge, D. and Kelly, J. P. (2006). Ways of computing diverse collaborative recommendations. In *Adaptive Hypermedia and Adaptive Web-based Systems*, pages 41–50.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370.

Carenini, G. (2001). GEA: a complete, modular system for generating evaluative arguments. In *International workshop on Computational Models of Natural Language Argument*, pages 959–968.

Carenini, G., Mittal, V., and Moore, J. (1994). Generating patient-specific interactive natural language explanations. *Proc Annu Symp Comput Appl Med Care*, pages 5–9.

Carenini, G. and Moore, D. J. (2001). An empirical study of the influence of user tailoring on evaluative argument effectiveness. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI 2001)*, pages 1307–1314.

Carenini, G. and Moore, J. (2000a). An empirical study of the influence of argument conciseness on argument effectiveness. In *ACL*, pages 150 – 157.

Carenini, G. and Moore, J. A. (2000b). A strategy for generating evaluative arguments. In *Proceedings of the 1st International Conference on Natural Language Generation*, pages 47–54.

Chen, L. and Pu, P. (2002). Trust building in recommender agents. In *WPRSIUI in conjunction with Intelligent User Interfaces*, pages 93–100.

Chen, L. and Pu, P. (2007). Hybrid critiquing-based recommender systems. In *Intelligent User Interfaces*, pages 22–31.

Cho, Y., Im, I., and Hiltz, J. F. S. R. (2003). The impact of product category on customer dissatisfaction in cyberspace. *Business Process Managment Journal*, 9 (5):635–651.

Clemen, R. T. (1996). *Making Hard Decisions*. Duxbury Press.

Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., and Riedl, J. (2003). Is seeing believing?: how recommender system interfaces affect users' opinions. In *CHI*, volume 1 of *Recommender systems and social computing*, pages 585–592.

Cramer, H., Evers, V., Someren, M. V., Ramlal, S., Rutledge, L., Stash, N., Aroyo, L., and Wielinga, B. (2008a). The effects of transparency on perceived and actual competence of a content-based recommender. In *Semantic Web User Interaction Workshop, CHI*.

Cramer, H. S. M., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., and Wielinga, B. J. (2008b). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Model. User-Adapt. Interact*, 18(5):455–496.

Cunningham, P., Doyle, D., and Loughrey, J. (2003). An evaluation of the usefulness of case-based explanation. In *In Proceedings of the Fifth International Conference on Case-Based Reasoning*, pages 122–130. Springer.

Czarkowski, M. (2006). *A Scrutable Adaptive Hypertext*. PhD thesis, University of Sydney.

Deerwester, S., Dumain, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 6:391–407.

Doyle, D., Tsymbal, A., and Cunningham, P. (2003). A review of explanation and explanation in case-based reasoning. Technical report, Department of Computer Science, Trinity College, Dublin.

Druzdzel, M. J. (1996). Qualitative verbal explanations in bayesian belief networks. *Artificial Intelligence and Simulation of Behaviour Quarterly, special issue on Bayesian networks*, pages 43–54.

Felfernig, A. and Gula, B. (2006). Consumer behavior in the interaction with knowledge-based recommender applications. In *ECAI 2006 Workshop on Recommender Systems*, pages 37–41.

Felfernig, A., Gula, B., G. Letiner, a. M. M., Melcher, R., Schippel, S., and Teppan, E. (2008). A dominance model for the calculation of decoy products in recommendation environments. In *AISB Symposium on Persuasive Technology*, pages 43–50.

Ferman, A. M., Errico, J. H., van Beek, P., and Sezan, M. I. (2002). Content-based filtering and personalization using structured metadata. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 393–393, New York, NY, USA. ACM Press.

Finch, I. (1998). Knowledge-based systems, viewpoints and the world wide web. In *IEE Colloquium on Web-Based Knowledge Servers*, pages 8/1–8/4.

Fogg, B., Marshall, J., Kameda, T., Solomon, J., Rangnekar, A., Boyd, J., and Brown, B. (2001). Web credibility research: A method for online experiments and early study results. In *CHI 2001*, pages 295–296.

Fogg, B. J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., and Tauber, E. R. (2003). How do users evaluate the credibility of web sites?: a study with over 2,500 participants. In *Proceedings of DUX'03: Designing for User Experiences*, number 15, pages 1–15.

Ginty, L. M. and Smyth, B. (2002). Comparison-based recommendation. *Lecture Notes in Computer Science*, 2416:731–737.

Good, N., Schafer, J. B., Konstan, J. A., Borchers, A., Sarwar, B. M., Herlocker, J. L., and Riedl, J. (1999). Combining collaborative filtering with personal agents for better recommendations. In *AAAI/IAAI*, pages 439–446.

Hance, E. and Buchanan, B. (1984). *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project*. Addison-Wesley.

Häubl, G. and Trifts, V. (2000). Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Science*, 19:4–21.

Herlocker, J. (2000). *Understanding and Improving Automated Collaborative Filtering Systems*. PhD thesis, University of Minnesota.

Herlocker, J., J., J. K., Borchers, A., and Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. In *Conference on Research and Development in Information Retrieval*.

Herlocker, J. L., Konstan, J. A., and Riedl, J. (2000). Explaining collaborative filtering recommendations. In *ACM conference on Computer supported cooperative work*, pages 241–250.

Herlocker, J. L., Konstan, J. A., Terveen, L., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53.

Hingston, M. (2006). User friendly recommender systems. Master's thesis, Sydney University.

Hofmann, T. (2003). Collaborative filtering via gaussian probabilistic latent semantic analysis. In *Annual ACM Conference on Research and Development in Information Retrieval - Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*.

Hu, Y., Koren, Y., and Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *ICDM*.

Hunt, J. E. and Price, C. J. (1988). Explaining qualitative diagnosis. *Engineering Applications of Artificial Intelligence*, 1(3):Pages 161–169.

Jacobs, C. E., Finkelstein, A., and Salesin, D. H. (1995). Fast multiresolution image querying. In *Proceedings of SIGGRAPH 95*, pages 277–286.

Joachims, T., Granka, L., and Pan, B. (2005). Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR conference on Research and development in information retrieval*, pages 154–161.

Jurafsky, D. and Martin, J. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing Computational Linguistics, and Speech Recognition*. Prentice Hall.

Khan, O. Z., Poupart, P., and Black, J. P. (2009). Minimal sufficient explanations for mdps. In *Workshop on Explanation-Aware Computing associated with IJCAI*.

Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In *EMNLP*, pages 423–430.

Kincaid, J. P., Jr., R. P. F., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas. Technical report, TN: Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

Krause, S. (2006). Pandora and last.fm: Nature vs. nurture in music recommenders. *http://www.stevekrause.org*.

Krulwich, B. (1997). The infofinder agent: Learning user interests through heuristic phrase extraction. *IEEE Intelligent Systems*, 12:22–27.

Laband, D. N. (1991). An objective measure of search versus experience goods. *Economic Inquiry*, 29 (3):497–509.

Lacave, C. and Diéz, F. J. (2002). A review of explanation methods for bayesian networks. *The Knowledge Engineering Review*, 17:2:107–127.

Lacave, C. and Diéz, F. J. (2004). A review of explanation methods for heuristic expert systems. *The Knowledge Engineering Review*, 17:2:133–146.

Lewis, D. D., Schapire, R. E., Callan, J. P., and Papka, R. (1996). Training algorithms for linear text classifiers. In *SIGIR*, pages 298–306.

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

Lopez-Suarez, A. and Kamel, M. (1994). Dykor: a method for generating the content of explanations in knowledge systems. *Knowledge-based Systems*, 7(3):177–188.

Masthoff, J. (2004). Group modeling: Selecting a sequence of television items to suit a group of viewers. *User Modeling and User Adapted Interaction*, 14:37–85.

Maybury, M., Greiff, W., Boykin, S., Ponte, J., McHenry, C., and Ferro, L. (2004). Personalcasting: Tailored broadcast news. *User Modeling and User-Adapted Interaction*, 14:119–144.

McCarthy, K., Reilly, J., McGinty, L., and Smyth, B. (2004). Thinking positively - explanatory feedback for conversational recommender systems. In *Proceedings of the*

European Conference on Case-Based Reasoning (ECCBR-04) Explanation Workshop, pages 115–124.

McCarthy, K., Reilly, J., Smyth, B., and Mcginty, L. (2005). Generating diverse compound critiques. *Artificial Intelligence Review*, 24:339–357.

McGinty, L. and Smyth, B. (2002). Comparison-based recommendation. *Lecture Notes in Computer Science*, 2416:575–589.

McNee, S., Lam, S.K.and Guetzlaff, C., and Konstan, J.A.and Riedl, J. (2003a). Confidence displays and training in recommender systems. In *INTERACT IFIP TC13 International Conference on Human-Computer Interaction*, pages 176–183.

McNee, S. M., Lam, S. K., Konstan, J. A., and Riedl, J. (2003b). Interfaces for eliciting new user preferences in recommender systems. *User Modeling*, pages pp. 178–187.

McNee, S. M., Riedl, J., and Konstan, J. A. (2006a). Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems (CHI 2006)*, pages 1097–1101.

McNee, S. M., Riedl, J., and Konstan, J. A. (2006b). Making recommendations better: An analytic model for human-recommender interaction. In *In the Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems (CHI 2006)*, pages 1103–1108.

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

McSherry, D. and Aha, D. W. (2007). Avoiding long and fruitless dialogues in critiquing. In *SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 173–186.

Mehta, B. (2007). Unsupervised shilling detection for collaborative filtering. In *AAAI*.

Melville, P., Mooney, R. J., and Nagarajan, R. (2002). Content-boosted collaborative filtering for improved recommendations. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence(AAAI-2002)*, pages 187–192.

Merlino, A., Morey, D., and Maybury, M. (1997). Broadcast news navigation using story segmentation. *ACM Multimedia*, pages pp. 381–391.

Mobasher, B., Burke, R., Williams, C., and Bhaumik, R. (2006). Analysis and detection of segment-focused attacks against collaborative recommendation. In *WebKDD Workshop*, pages 96–118.

Murphy, P. E. and Enis, B. M. (1986). Classifying products strategically. *Journal of Marketing*, 50:24–42.

Nguyen, H. and Masthoff, J. (2008). Using digital images to enhance the credibility of information. In *Persuasive Technology symposium in association with the Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB)*, pages 1–8.

Nielsen, J. and Molich, R. (1990). Heuristic evaluation of user interfaces. In *ACM CHI'90*, pages 249–256.

Ohanian, R. (1990). Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness. *Journal of Advertising*, 19:3:39–52.

O'Sullivan, D., Smyth, B., Wilson, D. C., McDonald, K., and Smeaton, A. (2004). Improving the quality of the personalized electronic program guide. *User Modeling and User-Adapted Interaction*, 14:5–36.

Paramythis, A., Totter, A., and Stephanidis, C. (2001). A modular approach to the evaluation of adaptive user interfaces. In Weibelzahl, S., Chin, D. N., and Weber, G., editors, *Evaluation of Adaptive Systems in conjunction with UM'01*, pages 9–24.

Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, 13:393–408.

Pazzani, M. J. (2002). Commercial applications of machine learning for personalized wireless portals. In *Pacific Rim Conference on Artificial Intelligence*, pages 1–5.

Pittarello, F. (2004). *Personalized digital television*, chapter 11, pages 287–320. Kluwer Academic Publishers.

Pu, P. and Chen, L. (2006). Trust building with explanation interfaces. In *IUI'06*, Recommendations I, pages 93–100.

Pu, P. and Chen, L. (2007). Trust-inspiring explanation interfaces for recommender systems. *Knowledge-based Systems*, 20:542–556.

Rafter, R. and Smyth, B. (2005). Conversational collaborative recommendation - an experimental analysis. *Artif. Intell. Rev*, 24(3-4):301–318.

Rashid, A. M., Albert, I., Cosley, D., Lam, S. K., McNee, S. M., Konstan, J. A., and Riedl, J. (2002). Getting to know you: learning new user preferences in recommender systems. In *International Conference on Intelligent User Interfaces*, pages 127–134.

Reilly, J., McCarthy, K., McGinty, L., and Smyth, B. (2004a). Dynamic critiquing. In Funk, P. and González-Calero, P. A., editors, *ECCBR*, volume 3155 of *Lecture Notes in Computer Science*, pages 763–777. Springer.

Reilly, J., McCarthy, K., McGinty, L., and Smyth, B. (2004b). Explaining compound critiquing. In *UK Workshop on Case-Based Reasoning (UKCBR-04)*, pages 12–20.

Reilly, J., McCarthy, K., McGinty, L., and Smyth, B. (2004c). Incremental critiquing. In *SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 143–151.

Reiter, E. and Dale., R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press.

Reiter, E., Robertson, R., and Osman, L. (2003). Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence Journal*, 144:41–58.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, pages 175–186.

Resnick, P. and Varian, H. (1997). Recommender systems. *Communications of the ACM*, 40(3):56–58.

Roth-Berghofer, T., Schulz, S., Leake, D. B., and Bahls, D. (2008). Workshop on explanation-aware computing. In *ECAI*.

Roth-Berghofer, T., Tintarev, N., and Leake, D. B. (2009). Workshop on explanation-aware computing. In *IJCAI*.

Salton, G. and McGil, M. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Shapiro, C. (1983). Optimal pricing of experience goods. *The Bell Journal of Economics*, 14 (2):497–507.

Sinha, R. and Swearingen, K. (2002). The role of transparency in recommender systems. In *Conference on Human Factors in Computing Systems*, pages 830–831.

Sørmo, F., Cassens, J., and Aamodt, A. (2005). Explanation in case-based reasoning perspectives and goals. *Artificial Intelligence Review*, 24(2):109 – 143.

Stiff, J. B. (1994). *Persuasive Communication*, chapter 5, pages 94–98. Guilford Press.

Swearingen, K. and Sinha, R. (2002). Interaction design for recommender systems. In *Designing Interactive Systems*, pages 25–28.

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293.

Symeonidis, P., Nanopoulos, A., and Manolopoulos, Y. (2007). Feature-weighted user model for recommender systems. In *User Modeling*, pages 97–106.

Symeonidis, P., Nanopoulos, A., and Manolopoulos, Y. (2008). Justified recommendations based on content and rating data. In *WebKDD Workshop on Web Mining and Web Usage Analysis*.

Takács, G., Pilászy, I., Németh, B., and Tikk, D. (2007). Major components of the gravity recommendation system. *SIGKDD Explor. Newsl.*, 9(2):80–83.

Tanaka-Ishii, K. and Frank, I. (2000). Multi-agent explanation strategies in real-time domains. In *38th Annual Meeting on Association for Computational Linguistics*, pages 158–165.

Thompson, C. A., Göker, M. H., and Langley, P. (2004). A personalized system for conversational recommendations. *J. Artif. Intell. Res. (JAIR)*, 21:393–428.

Tintarev, N. and Masthoff, J. (2007). A survey of explanations in recommender systems. In *WPRSIUI associated with ICDE'07*, pages 801–810.

Tintarev, N. and Masthoff, J. (2008a). Over- and underestimation in different product domains. In *Workshop on Recommender Systems associated with ECAI*, pages 14–19.

Tintarev, N. and Masthoff, J. (2008b). Personalizing movie explanations using commercial meta-data. In *Adaptive Hypermedia*, pages 204–213.

Towle, B. and Quinn, C. (2000). Knowledge based recommender systems using explicit user models. In *Knowledge-Based Electronic Markets, Papers from the AAAI Workshop, AAAI Technical Report WS-00-04*, pages pp. 74–77, Menlo Park, CA. AAAI Press.

Vig, J., Sen, S., and Riedl, J. (2009). Tagsplanations: Explaining recommendations using tags. In *Intelligent User Interfaces*.

Wang, W. and Benbasat, I. (2007). Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Managment Information Systems*, 23:217–246.

Wärnestål, P. (2005a). Modeling a dialogue strategy for personalized movie recommendations. In *Beyond Personalization Workshop*, pages 77–82.

Wärnestål, P. (2005b). User evaluation of a conversational recommender system. In *Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 32–39.

Wick, M. R. and Thompson, W. B. (1992). Reconstructive expert system explanation. *Artif. Intell.*, 54(1-2):33–70.

Williams, S. and Reiter, E. (2008). Generating basic skills reports for low-skilled readers. *Journal of Natural Language Engineering*, 14(4):495–525.

Wolverton, M. (1995). Presenting significant information in expert system explanation. *Lecture Notes in Computer Science*, 990:435–440.

Ye, L., Johnson, P., Ye, L. R., and Johnson, P. E. (1995). The impact of explanation facilities on user acceptance of expert systems advice. *MIS Quarterly*, 19(2):157–172.

Yee, K.-P., Swearingen, K., Li, K., and Hearst, M. (2003). Faceted metadata for image search and browsing. In *ACM Conference on Computer-Human Interaction*, pages 401–408.

Zaslow, J. (2002). Oh no! My TiVo thinks I'm gay. *The Wall Street Journal*.

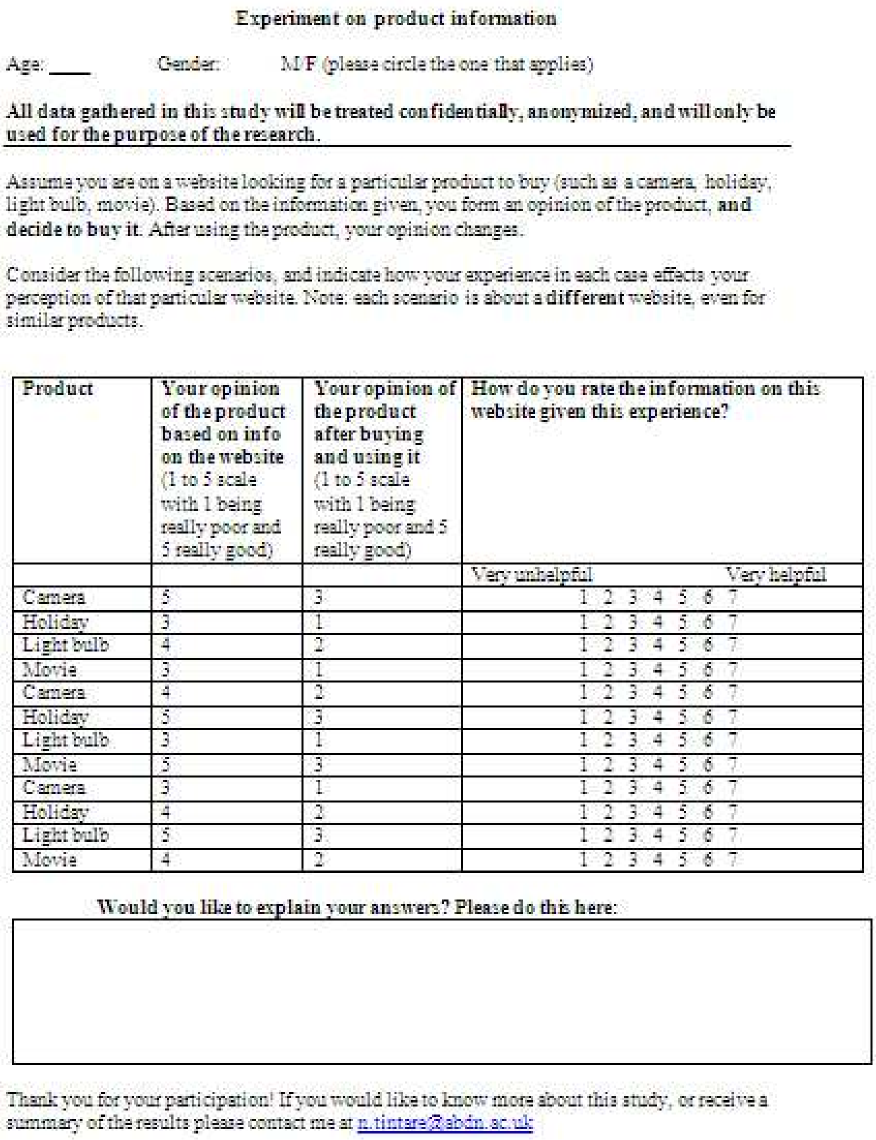
Ziegler, C.-N., McNee, S. M., Konstan, J. A., and Lausen, G. (2005). Improving recommendation lists through topic diversification. In *WWW'05*, pages 22–32.

Zimmerman, J., Kurapati, K., Buczak, A. L., Schaffer, D., Gutta, S., and Martino, J. (2004). *Personalized digital television*, chapter 2, pages 27–51. Kluwer Academic Publishers.

Appendix A

Questionnaires for perceived effectiveness

This appendix contains two example questionnaires for over- and underestimation. Note that there were eight versions of the questionnaire in total, for different orderings of domains and values.



Experiment on product information

Age: ____ Gender: M/F (please circle the one that applies)

All data gathered in this study will be treated confidentially, anonymized, and will only be used for the purpose of the research.

Assume you are on a website looking for a particular product to buy (such as a camera, holiday, light bulb, movie). Based on the information given, you form an opinion of the product, **and decide to buy it**. After using the product, your opinion changes.

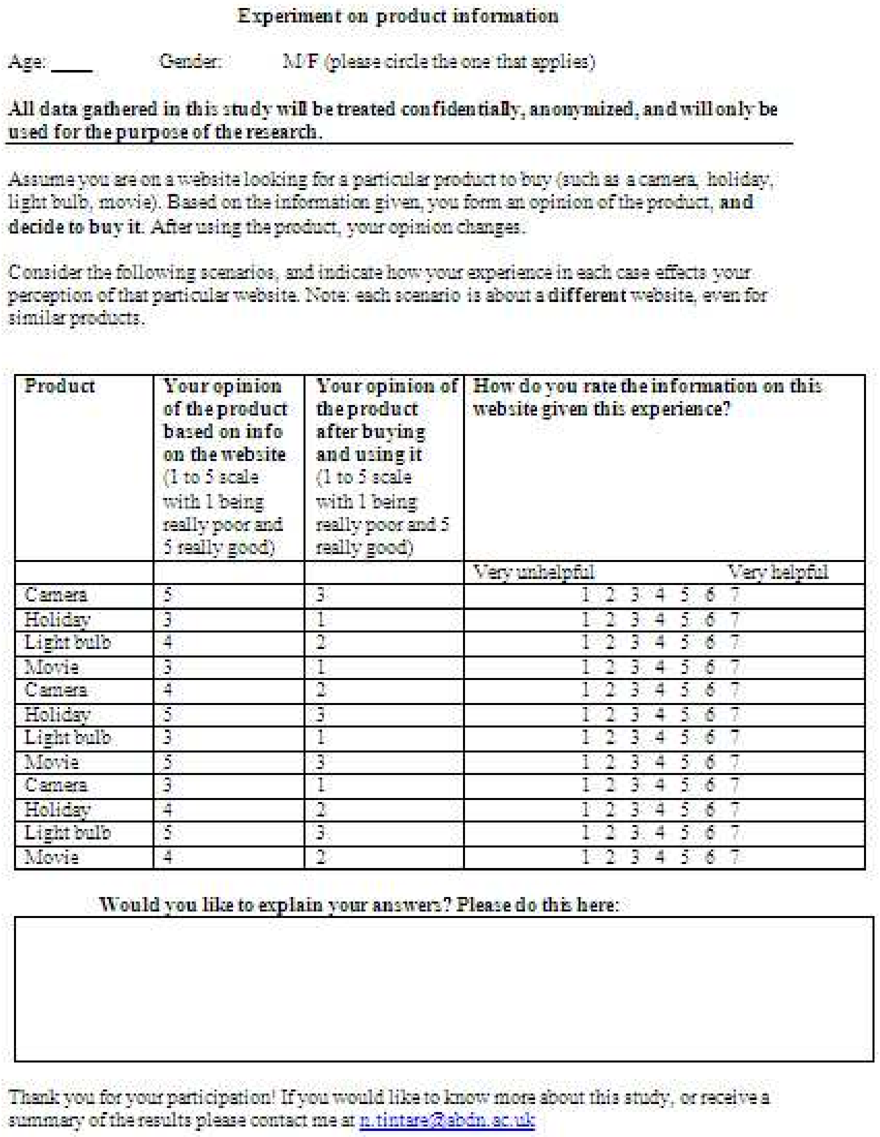
Consider the following scenarios, and indicate how your experience in each case effects your perception of that particular website. Note: each scenario is about a **different** website, even for similar products.

Product	**Your opinion of the product based on info on the website** (1 to 5 scale with 1 being really poor and 5 really good)	**Your opinion of the product after buying and using it** (1 to 5 scale with 1 being really poor and 5 really good)	**How do you rate the information on this website given this experience?**
			Very unhelpful Very helpful
Camera	5	3	1 2 3 4 5 6 7
Holiday	3	1	1 2 3 4 5 6 7
Light bulb	4	2	1 2 3 4 5 6 7
Movie	3	1	1 2 3 4 5 6 7
Camera	4	2	1 2 3 4 5 6 7
Holiday	5	3	1 2 3 4 5 6 7
Light bulb	3	1	1 2 3 4 5 6 7
Movie	5	3	1 2 3 4 5 6 7
Camera	3	1	1 2 3 4 5 6 7
Holiday	4	2	1 2 3 4 5 6 7
Light bulb	5	3	1 2 3 4 5 6 7
Movie	4	2	1 2 3 4 5 6 7

Would you like to explain your answers? Please do this here:

Thank you for your participation! If you would like to know more about this study, or receive a summary of the results please contact me at n.tintare@abdn.ac.uk

Figure A.1: Experiment on product information - Overestimation



Experiment on product information

Age: ____ Gender: M/F (please circle the one that applies)

All data gathered in this study will be treated confidentially, anonymized, and will only be used for the purpose of the research.

Assume you are on a website looking for a particular product to buy (such as a camera, holiday, light bulb, movie). Based on the information given, you form an opinion of the product, **and decide to buy it**. After using the product, your opinion changes.

Consider the following scenarios, and indicate how your experience in each case affects your perception of that particular website. Note: each scenario is about a **different** website, even for similar products.

Product	**Your opinion of the product based on info on the website** (1 to 5 scale with 1 being really poor and 5 really good)	**Your opinion of the product after buying and using it** (1 to 5 scale with 1 being really poor and 5 really good)	**How do you rate the information on this website given this experience?**
			Very unhelpful Very helpful
Camera	5	3	1 2 3 4 5 6 7
Holiday	3	1	1 2 3 4 5 6 7
Light bulb	4	2	1 2 3 4 5 6 7
Movie	3	1	1 2 3 4 5 6 7
Camera	4	2	1 2 3 4 5 6 7
Holiday	5	3	1 2 3 4 5 6 7
Light bulb	3	1	1 2 3 4 5 6 7
Movie	5	3	1 2 3 4 5 6 7
Camera	3	1	1 2 3 4 5 6 7
Holiday	4	2	1 2 3 4 5 6 7
Light bulb	5	3	1 2 3 4 5 6 7
Movie	4	2	1 2 3 4 5 6 7

Would you like to explain your answers? Please do this here:

Thank you for your participation! If you would like to know more about this study, or receive a summary of the results please contact me at n.tintare@abdn.ac.uk

Figure A.2: Experiment on product information - Underestimation

Appendix B

Content for movies

B.1 Focus groups

B.1.1 List of movies

Action/Adventure	Batman	Star Wars	Terminator 2
Animation	Toy Story	Aladdin	Beauty and the Beast
Children	Toy Story	Babe	Beauty and the Beast
Comedy	Back To The Future	Ace Ventura: pet detective	Mrs. Doubtfire
Crime/Gangster	Pulp Fiction	Godfather	The Silence of the Lambs
Documentary	March of the Penguins	The Fog of War	Bowling for Columbine
Drama	American Beauty	Forrest Gump	Shawshank Redemption
Fantasy	Batman	The Lord of the Rings: The Return of the King (III)	Toy Story
Film Noir	Sunset Boulevard	Double Indemnity	The Maltese Falcon
Epics/Historical	Schindler's List	Hotel Rwanda	Braveheart
Horror	Jaws	Psycho	Alien
Musicals	The Wizard of Oz	Beauty and the Beast	Singin' in the Rain
Mystery	Memento	The Usual Suspects	Citizen Kane
Romance	Forrest Gump	Beauty and the Beast	Pretty Woman
Science Fiction	Star Wars	Terminator 2	Back To The Future
Thriller	Batman	The Silence of the Lambs	Terminator 2
War	Schindler's List	Lawrence of Arabia	Apocalypse Now
Western	Butch Cassidy and the Sundance Kid	The Good, the Bad and the Ugly	Once Upon a Time in the West

B.1.2 Participant introductions

This section describes participant introductions, describing their favorite movie and justification for this movie. Below **"P"** denotes a participant, and **"F"** the facilitator.

1. P:"I have no idea um, oh yeah, just because we mentioned it with [name] yesterday. Star wars is unbelievable uh, just because uh it's so unbelievable. They invented special effects which were beautiful. It's just a great story about phh human drama, of good and bad. And princess Leia is so sexy!"

2. P:"Uhh . . . I like uh Bladerunner. It's good as action, the science fiction, and there's

some deeper undercurrents sometimes, well, you can just take the action."

3. P:"Ok, my favourite movie is the Sound of Music, because I watched it all the time when I was a kid, and I know all the words all by heart."

4. P:"My favourite movie has to be the Matrix, the one, because I thought the idea was like incredible and it's the kind of movie that, the first time you see it, you don't understand like everything, but then like when you watch it several times like you every time like discover something new. And the actions scenes are like very good as well, yeah that's my favourite."

5. P:"One of my favourites is [cinema] Paradiso. It's just kind of sad, but it's really good."
   F: "So, why did you say you liked it?"
   P: "It's kind of sad, but it's just a really nice story I think. I don't necessarily like sad films, but I just like that one."

6. P:"For me, I like also the starwars, but I like also the Indiana jones, but I think maybe the point that I saw them when we was very young so I have a very good remembering of the first watching. So now like it's the best because they were the best around."

7. P:"Ok, so we start with Jack Nicholson as a string of, well, guidance. Let's say, phh, Chinatown? Has anyone seen Chinatown? It's like Los Angeles in the 50s or something like that. He's playing like some noir detective and so um there is like some corrupted politicians all around and some weird mafia stories or something like that. And basically there is some woman which is trying to run away [F: just a few sentences why you like it]..um well I like it because its really really bad for him at the end of the movie, he suffers a lot, a lot, they are like um it's really well made they are really sadistic, I like it."

8. P:"Yeah I don't know, I don't think I have a favourite movies just depends on the mood what I like to watch, but I like martial arts movies in general, like Jet Li, so. I don't, normally I don't have a favourite actor or actress, but Jet Li is probably one of my favourite actors, but anything from him is good: Hero uh was the most recent movie, it was good. Follows [Participant1] topic, recently Crash was very good, uh the scene was set in LA as well, talking about corrupted police men, racism and all sort of things."

9. P:"Ehm, yep, favorite movies would be, ehm, Good Fellas or Casino. Ehm, I quite like the old gangster genre, ehm, especially like eh, Martin Scorsese as well as a director eh especially like films like good fellas and casino because it's real life

stories as well"

F: "So you find it realistic?"
P: "A bit of realism. Ehm yeah, sometimes I'm not too keen on the ehm kind of the stereotypical Hollywood endings and things like that. Usual suspects is another good one, ehm, which always keeps you guessing"

10. P:"Ehm, probably Withnail and I, I mean there is a few others that are good. But just Withnail and I - funny, funny, funny until the last bit, which just, the last bit just puts everything into perspective, just something very powerful about it, but yeah, also like, yeah, no real favourite movie really, that's just off the top of my head."

    F: "Which one did you say?"

    P:"Withnail and I, Staring a very young Richard Grant"

11. P: "I think it's Red Lines of Terrence Malick. Eh it's a films about the war, and ehm it's um very poetic, and ehm, but the subject is very strong, but ehm, there is a, it's not about one actor or one characters, it's eh of many characters who were quite forgot in the war."

B.2 Features vs. detail (part1)

Voluntary study about movie reviews Part 1/2:

Gender: M/F **Age:** _ _ _ _

Instructions:

Pretend that it's Sunday afternoon, and you want to choose a movie to rent and watch at home with a close friend.

You are alone in the rental shop and pick up this title. Your friend has similar tastes to you, and is not too picky anyway. You wonder if this one is worth watching, or if you should keep looking, and see if there is anything better. To help decide you will be shown two reviews of this movie.

Review A:

This movie is at once a moving drama, and a chilling thriller. The plot is about a low-level British diplomat who has always gone about his work very quietly, not causing any problems along the way. But after the murder of his wife, he feels compelled to find out why, and is thrust into the middle of an ugly conspiracy. Ample justice is done to the beauty of the African continent. This movie shows how pharmaceutical companies affect the lively but trampled-upon people of Kenya.

Review B:

This movie is at once a moving drama, and a chilling thriller. The plot is about a low-level British diplomat who has always gone about his work very quietly, not causing any problems along the way. But after the murder of his wife, he feels compelled to find out why, and is thrust into the middle of an ugly conspiracy. [Name1], director of [Another movie], has made yet another gem of a movie. The quality of Africa, of Kenya, and of the African people is recreated by the shots of [Name2]. This movie shows how pharmaceutical companies affect the lively but trampled-upon people of Kenya.

Which review do you think is better, A or B?

Mark your preference on the scale below.

Definitely A **(same)** **Definitely B**

□ □ □ □ □ □ □

Why?

□

B.3 Which features do you find important? (part 2)

Below we list possible features that can be used to describe a movie. If you are unsure what we mean by a certain feature, read the example quotes for reference. **Please tick up to 5 features you think should be mentioned in a movie review, in addition to a summary.**

Feature	**Example**	**Important?**
Cast	*David Hayman's performance as Boyle in this film is powerful and subtle.*	
Director	*Terry Gilliam is a director who makes interesting films.*	
Dialogs	*... the one-liners are hilarious*	
Good in its genre	*... the drama is tense and sometimes unbearable.*	
Group/alone	*... it's best to watch alone or with a lover ...*	
Good for kids	*A charming family film which everyone can enjoy.*	
Initial expectations	*A fairly faithful adaptation of Jimmy Boyle's autobiography*	
Suites mood	*... a real feel-good movie*	
Movie Studio	*... yet another gem from Disney*	
Originality	*... it's not the predictable Hollywood crap*	
Pace	*The movie takes its time ...*	
Realistic	*The animated animals could have been real!*	
Easy viewing	*An intriguing brainteaser ...*	
Repulsive/violent	*... provided you don't mind the fact that there are many violent scenes*	
Sex	*... a puerile adolescent sex movie ...*	
Soundtrack	*The outstanding soundtrack by ...*	
Subject matter	*... an interesting insight into a very small part of the Rwandan civil war ...*	
Visuals (incl. special effects, animations)	*... with Ang Lee's amazing cinematography.*	

B.4 Review texts used

Summary (same in all):

The plot is about a low-level British diplomat who has always gone about his work very quietly, not causing any problems along the way. But after the murder of his wife, he feels compelled to find out why, and is thrust into the middle of an ugly conspiracy.

No detail, 4 features

This movie is a drama and a thriller. **[Summary]** The directing makes this a fabulous movie. The photography is beautiful. The African continent and the African people are vividly portrayed.

Detail, 2 features

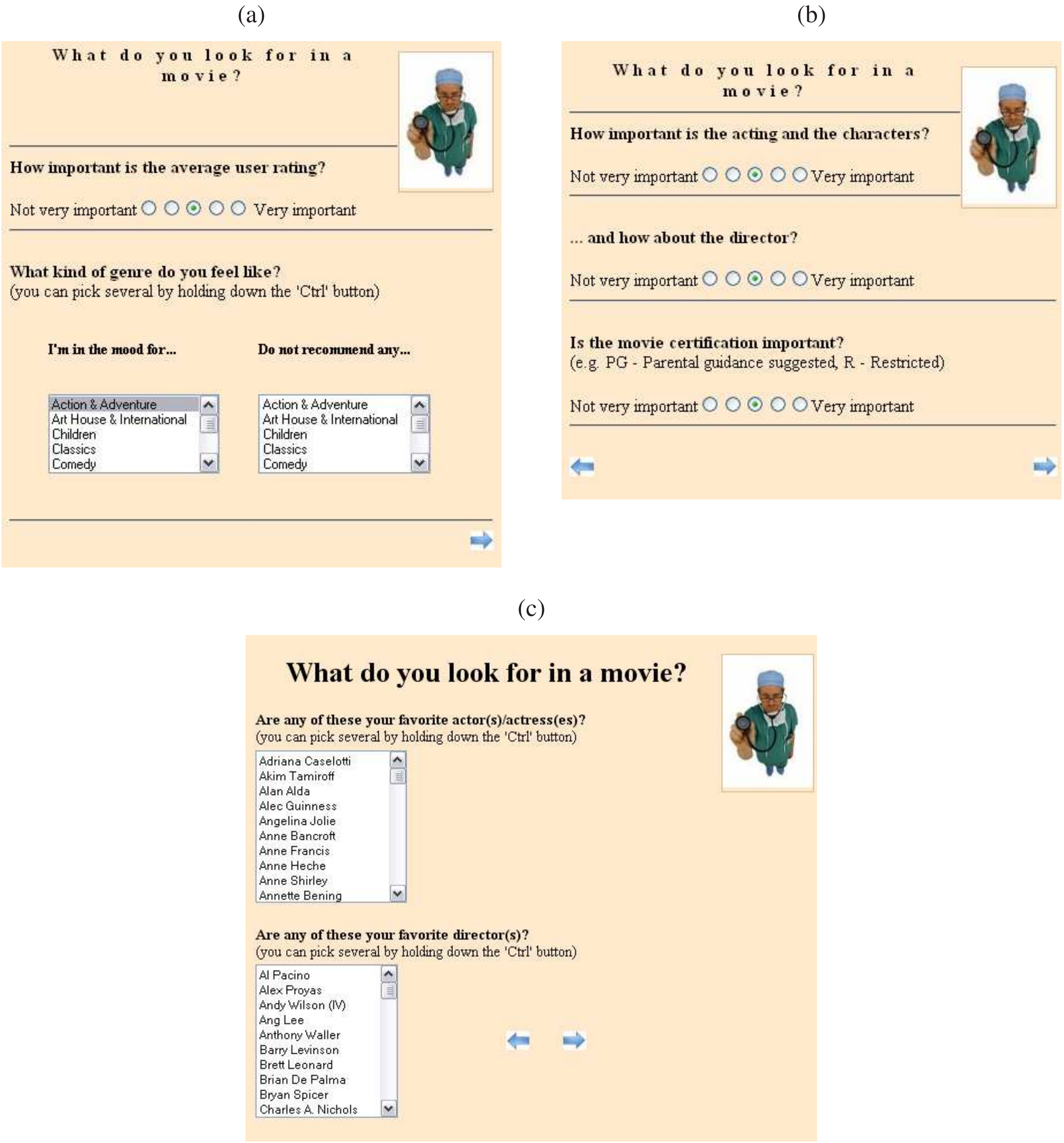
This movie is at once a moving drama, and a chilling thriller. **[Summary]** Ample justice is done to the beauty of the African continent. This movie shows how pharmaceutical companies affect the lively but trampled-upon people of Kenya.

Detail, 4 features

This movie is at once a moving drama, and a chilling thriller. **[Summary] [Name1]**, director of **[Another movie]**, has made yet another gem of a movie. The quality of Africa, of Kenya, and of the African people is recreated by the shots of **[Name2]**. This movie shows how pharmaceutical companies affect the lively but trampled-upon people of Kenya.

B.5 Screenshots Movies

Figure B.1: Input of user preferences



(a)

What do you look for in a movie?

How important is the average user rating?

Not very important ○ ○ ● ○ ○ Very important

What kind of genre do you feel like?
(you can pick several by holding down the 'Ctrl' button)

I'm in the mood for...	Do not recommend any...
Action & Adventure Art House & International Children Classics Comedy	Action & Adventure Art House & International Children Classics Comedy

(b)

What do you look for in a movie?

How important is the acting and the characters?

Not very important ○ ○ ● ○ ○ Very important

... and how about the director?

Not very important ○ ○ ● ○ ○ Very important

Is the movie certification important?
(e.g. PG - Parental guidance suggested, R - Restricted)

Not very important ○ ○ ● ○ ○ Very important

(c)

What do you look for in a movie?

Are any of these your favorite actor(s)/actress(es)?
(you can pick several by holding down the 'Ctrl' button)

Adriana Caselotti
Akim Tamiroff
Alan Alda
Alec Guinness
Angelina Jolie
Anne Bancroft
Anne Francis
Anne Heche
Anne Shirley
Annette Bening

Are any of these your favorite director(s)?
(you can pick several by holding down the 'Ctrl' button)

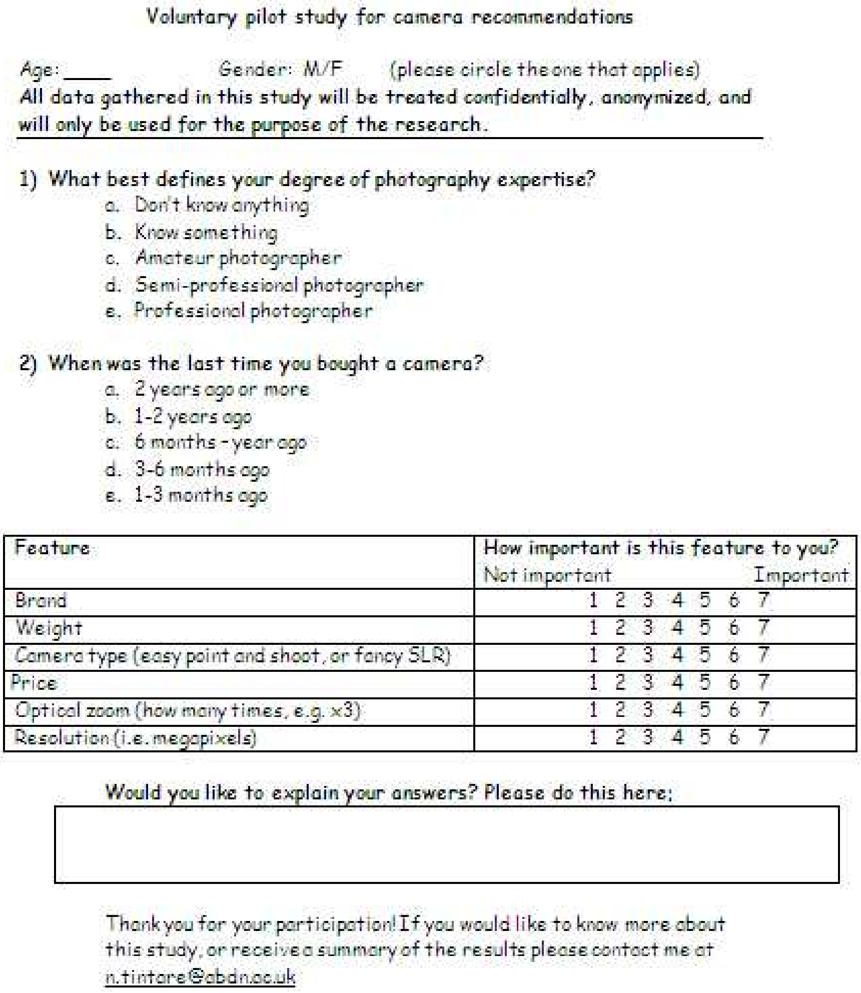
Al Pacino
Alex Proyas
Andy Wilson (IV)
Ang Lee
Anthony Waller
Barry Levinson
Brett Leonard
Brian De Palma
Bryan Spicer
Charles A. Nichols

B.6 Cameras

This appendix contains the questionnaire for eliciting important camera features, and screenshots from the camera explanation experiment.

B.6.1 Camera features questionnaire

Figure B.2: Voluntary pilot study for camera recommendations



Voluntary pilot study for camera recommendations

Age: ____ Gender: M/F (please circle the one that applies)

All data gathered in this study will be treated confidentially, anonymized, and will only be used for the purpose of the research.

1) What best defines your degree of photography expertise?

a. Don't know anything
b. Know something
c. Amateur photographer
d. Semi-professional photographer
e. Professional photographer

2) When was the last time you bought a camera?

a. 2 years ago or more
b. 1-2 years ago
c. 6 months - year ago
d. 3-6 months ago
e. 1-3 months ago

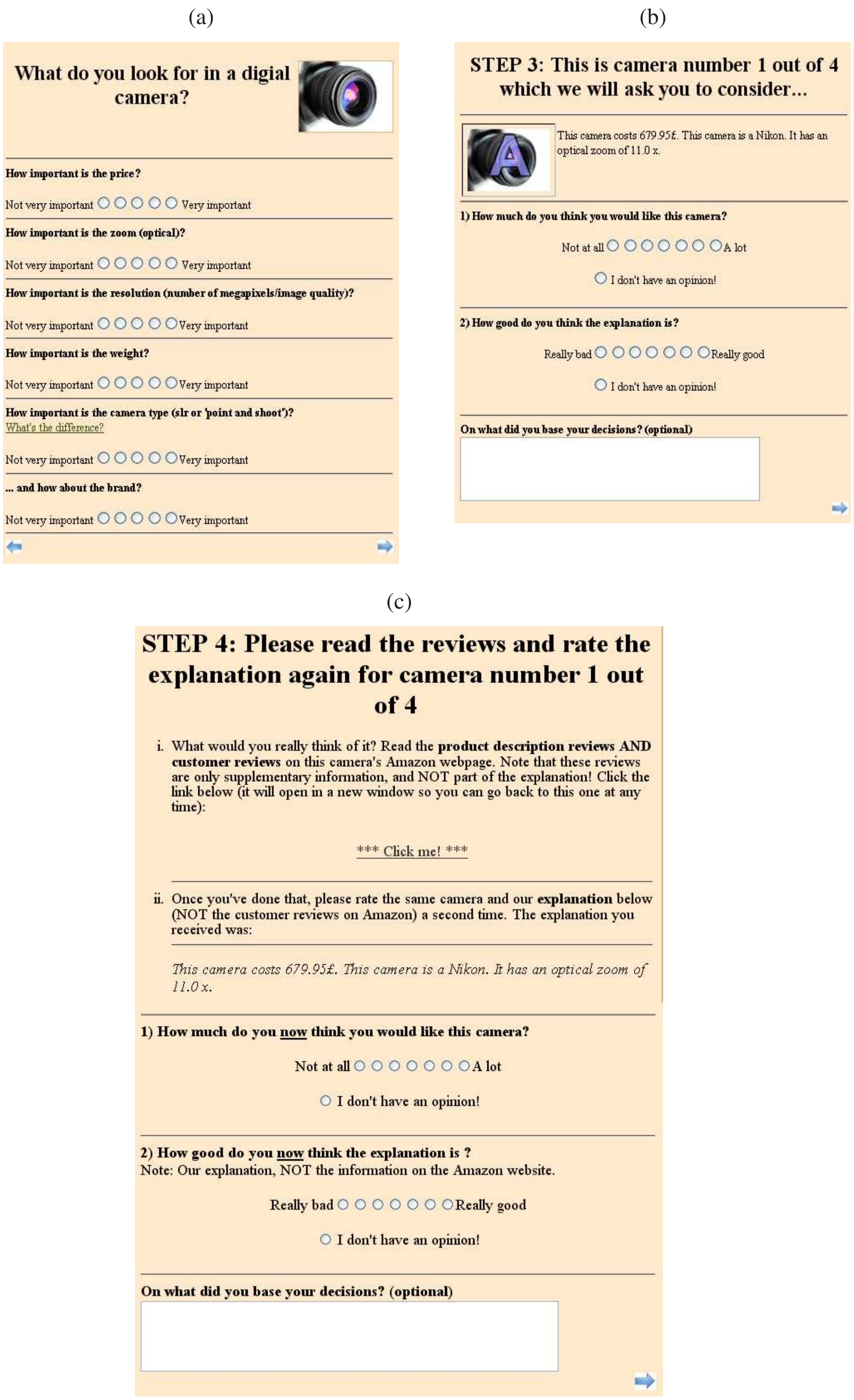
Feature	How important is this feature to you? Not important Important
Brand	1 2 3 4 5 6 7
Weight	1 2 3 4 5 6 7
Camera type (easy point and shoot, or fancy SLR)	1 2 3 4 5 6 7
Price	1 2 3 4 5 6 7
Optical zoom (how many times, e.g. x3)	1 2 3 4 5 6 7
Resolution (i.e. megapixels)	1 2 3 4 5 6 7

Would you like to explain your answers? Please do this here;

Thank you for your participation! If you would like to know more about this study, or receive a summary of the results please contact me at n.tintore@abdn.ac.uk

B.6.2 Screenshots

Figure B.3: Input of user preferences



(a)

What do you look for in a digial camera?

How important is the price?

Not very important ○ ○ ○ ○ ○ Very important

How important is the zoom (optical)?

Not very important ○ ○ ○ ○ ○ Very important

How important is the resolution (number of megapixels/image quality)?

Not very important ○ ○ ○ ○ ○Very important

How important is the weight?

Not very important ○ ○ ○ ○ ○Very important

How important is the camera type (slr or 'point and shoot')?
What's the difference?

Not very important ○ ○ ○ ○ ○Very important

... and how about the brand?

Not very important ○ ○ ○ ○ ○Very important

(b)

STEP 3: This is camera number 1 out of 4 which we will ask you to consider...

This camera costs 679.95£. This camera is a Nikon. It has an optical zoom of 11.0 x.

1) How much do you think you would like this camera?

Not at all ○ ○ ○ ○ ○ ○ ○A lot

○ I don't have an opinion!

2) How good do you think the explanation is?

Really bad ○ ○ ○ ○ ○ ○ ○Really good

○ I don't have an opinion!

On what did you base your decisions? (optional)

(c)

STEP 4: Please read the reviews and rate the explanation again for camera number 1 out of 4

i. What would you really think of it? Read the **product description reviews AND customer reviews** on this camera's Amazon webpage. Note that these reviews are only supplementary information, and NOT part of the explanation! Click the link below (it will open in a new window so you can go back to this one at any time):

*** Click me! ***

ii. Once you've done that, please rate the same camera and our **explanation** below (NOT the customer reviews on Amazon) a second time. The explanation you received was:

This camera costs 679.95£. This camera is a Nikon. It has an optical zoom of 11.0 x.

1) How much do you now think you would like this camera?

Not at all ○ ○ ○ ○ ○ ○ ○A lot

○ I don't have an opinion!

2) How good do you now think the explanation is ?
Note: Our explanation, NOT the information on the Amazon website.

Really bad ○ ○ ○ ○ ○ ○ ○Really good

○ I don't have an opinion!

On what did you base your decisions? (optional)

Appendix C

Testbed Implementation

C.1 Overview

This chapter describes the implementation of the explanation generation testbed implemented in Java under Eclipse and Netbeans. Figure C.1 outlines the architecture of the testbed, which has four components numbered as follows:

1. Extracting data from Amazon Webservices and inserting it into a local database (described in Section C.2).
2. Input of user preferences, and deduction of a simple user model (described in Section C.3).
3. Selecting items to recommend based on the available options in the database (described in Section C.4).
4. Generating an explanation for each selected item (described in Section C.5).

We will use a running example throughout this chapter to illustrate how the testbed may work. For the sake of simplicity only one example is given, in the movie domain.

C.2 Extracting the data from Amazon

C.2.1 Why Amazon?

For movies it would arguably have been better to use a webservice that provides richer information about the movies, such as the Internet Movie Database (IMDB)[1]. Likewise, a

[1] http://www.imdb.com, retrieved January 2009

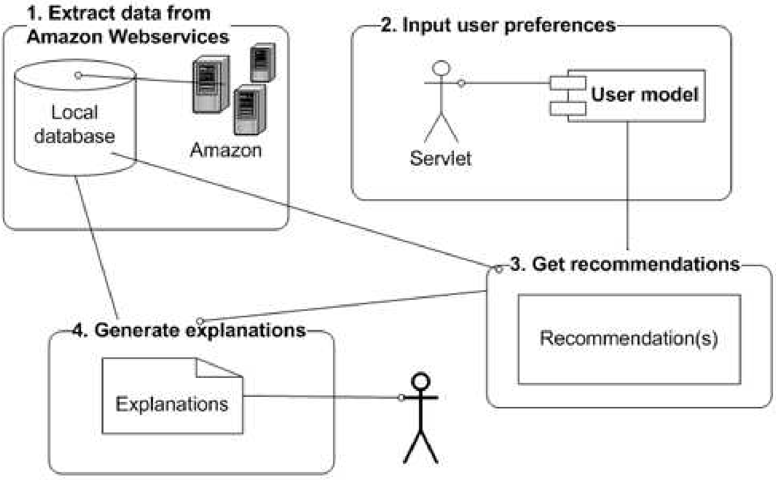


Figure C.1: Overview of implementation

better data source could no doubt be found for cameras. Our aim however, was to create a testbed that could be used in a number of domains. Amazon Webservices offered us a lot of flexibility in this regard. We also wanted to use data that is available and currently used (in some capacity) in a recommender system. Given that Amazon has a recommendation facility, and the Webservices are free, it felt like the most suitable choice.

C.2.2 Amazon Webservices

Amazon has a web service through which item features can be retrieved [2]. The same interface can be used for a variety of domains, making this testbed widely applicable.

Information about items can be retrieved via both SOAP (Simple Object Access Protocol) and REST (Representative State Transfer) interfaces. We elected to use the SOAP interface, because the code required for interaction with the webservices can be automatically generated (more about this in Section C.2.3). While the Amazon service was free of charge it did require registration, and any retrieval required reference to an access key id. Registration also came with terms of service, such as restrictions on how long retrieved data could be stored in a local database or cache. While these terms are likely to be modified over time, any developer using these services would be strongly advised to read the terms of service.

[2] When we first used this interface in 2006 this interface was called e-Commerce Service (ECS), but has since changed name to Amazon Webservices

C.2.3 Eclipse + Xfire plugin

As previously mentioned, we used a SOAP interface to retrieve data from the Amazon Webservice. SOAP works in a way similar to remote procedure calls (RPCs), but also wraps all communication in extra packaging, called an envelope. For such a small application, using SOAP only makes sense if the stubs for the remote procedure calls are automatically generated. Otherwise generating all the relevant functions would be a very time consuming practice. Once the stubs are generated the developer calls a function locally which triggers the retrieval of a value or data type from a remote interface, even if it is nested deep into an XML tree.

However, for stubs to be generated a description file needs to be present. Amazon supplies such Web-Service Definition Files (or WSDLs) for their services. These files are very large and describe an entire communication protocol, that is all of the information necessary for a client to invoke the methods of a web service: the data types used as method parameters or return values, individual methods names and signatures (WSDL refers to methods as operations), protocols and message formats allowed for each method, and the URLs used to access the web service. We also note that Amazon supplies a number of different definition files depending on location (i.e. the structure of data differs between countries). The Eclipse plugin XFire [3] then uses this description file to automatically generate stubs.

A practical tutorial for getting started with Amazon ECS using Eclipse and Xfire [4], and our more elaborate tutorial which also describes common problems [5] are both available online.

C.2.4 Look up by item (ItemLookup, ItemSearch)

Looking up an item in Amazon can be done in a number of ways. Firstly, each item has a unique identifier called ASIN. If you know the ASIN, you can look up the properties of the item using the ItemLookUp class. Note however that ASINs differ between countries, i.e. the same ASIN does not identify the same item on Amazon.co.uk as Amazon.com.

Items can also be looked up using a search class (ItemSearch) whereby Amazon returns items with similar names. You can at the same time restrict the search, e.g. by limiting it to a certain domain or “search index” such as DVDs. This list can be rather long, split up in pages of 10 items. Also, these are sorted by name by default. One way of creating variation is selecting e.g. the first item on each page rather than choosing each consecutive item on a page.

[3] http://xfire.codehaus.org/, retrieved January 2009

[4] http://xfire.codehaus.org/Eclipse+Plugin, retrieved January 2009

[5] http://www.csd.abdn.ac.uk/˜ ntintare/Tutorial1f.htm, retrieved January 2009

C.2.5 Look up by genre and Browsenodes

The word genre here is chosen to denote a hierarchical classification, although this word is particularly suitable for classification in the movie domain. Amazon webservices represent data as (upside-down) trees of so called "Browsenodes". In this way, one can retrieve subtrees of items which are classified under that Browsenode. This structure is useful if you know which classification you require (e.g. Thriller), but not what specific items. Each Browsenode is identified by a unique identifier, which can be looked up online[6]. Similarly to ASINs, Browsenode ids differ between countries. Also, the categories are randomly assigned and re-assigned. This means that similar id numbers do not necessarily imply similar categories, and that an id may change over time.

Analogous to this, if you want to find out what genres a found movie belongs to you also use Browsenodes. This method has a few faults however. Traversal of Browsenodes moves upward (from children to parents), and an item can have multiple parents. When a node has more than one parent node, the BrowseNodes response group only returns data for one of the parents. There is no logic that determines which of the parent nodes it follows up the ancestral tree. Running the request multiple times, therefore might return a different set of ancestors for a node. So if a movie belongs to multiple genres, there is no guarantee which one you will identify first. This can be remedied by iterating over all parents of a Browsenode.

Another limitation occurs if you are searching for items by BrowseNode: there is no guarantee that the retrieved items are of type you were looking for, e.g. you might reach an item that is not a movie. This can be remedied by stopping the search once a genre is found (this also decreases search time when surveying all parents nodes). For this you need to know what genres are valid BrowseNodes however.

C.2.6 Look up by similarity (SimilarityLookup)

While this is not a feature we have taken advantage of, it is one that could be very useful for explanation generation. Given an item, Amazom Webservices can return a list of similar items. The number of returned items is however limited to 10 per lookup. To get a longer list of similar items it possible to call SimilarityLookup several times. It is however non-deterministic, that is, each time the call is made different items may be returned. A developer would therefore need to check for duplicates, there is also no way to control for domain as there may be similar versions of the same item in different domains i.e. a book can be considered similar to a movie. Domain can be controlled as genres described in Section C.2.5. One way to make sure no duplicates are recorded in Java is to store items

[6] http://www.browsenodes.co.uk/ for Amazon.co.uk and http://www.browsenodes.com/ for Amazon.com, retrieved January 2009

as a set.

The retrieved similar items can consequently be used when generating explanations. In the simplest form, the explanation can simply list the items. Amazon uses correlations between users in terms of purchase habits, so this explanation might be collaborative-based, and read along the lines of: “People who bought this item also bought items X,Y and Z”. It would also be possible to make a more in-depth comparison between items on a feature level, in order to generate content-based explanations. For example, these explanations might read something like: “You might like Movie A, because it also stars Actor B”.

C.2.7 Database

The retrieved data can be stored in a database of the developers choosing as long as the terms of service are adhered to with regard to factors such as frequency of update. An access database was used initially as a proof of concept, and was later replaced by a MySQL database. When populating the database it is important to recognize that item features can be null. All properties exist for all items, regardless of domains, but are null in irrelevant domains. So, while you can look up optical zoom for movies in a similar way you will look up actors the value will of course be null. Null values may also occur in the appropriate domains (e.g. some cameras do not have optical zoom specified), and need to be checked.

Movies

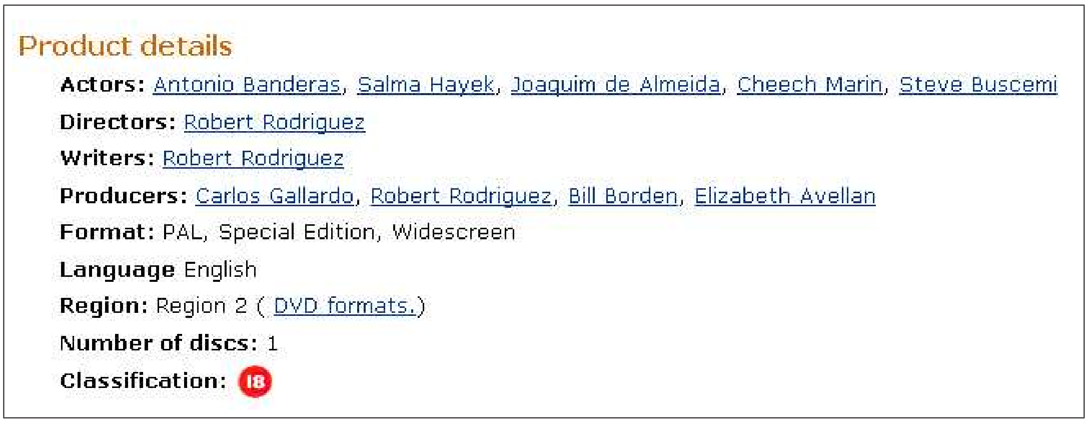
For movies we store the following features: genre, actors, directors, MPAA ratings, and average rating. Table C.1 shows an example movie object. These features were a compromise between the features available on Amazon and the ones that were mentioned in our user studies in Chapter 5. While genres are extracted via Browsenodes, the remaining features can be retrieved as “item attributes” for each movie. Image C.2 shows an example of the types of features that are available via Amazon for movies.

Cameras

For cameras we stored the following features: brand, image resolution, optical zoom, price, type and weight. We had similar problems using Browsenodes for cameras as we did for movies. That is, we had to ensure that the returned items were cameras and not peripheral items such as lenses. Given time constraints, and the small number of cameras required, we chose to insert the cameras by hand. In addition, this allowed us to ensure the features were well distributed across cameras (e.g. similar numbers of each brand, wide price range) as well as making sure there were at least three reviews. Image C.3

Table C.1: Example movie object

ASIN	B0000A2ZU1
Title	Desperado [1996]
URL	http://www.amazon.com. . .
Image URL	http://www.amazon.com. . .
Top 250	False
Genres	Action & Adventure, Children
Actors	Antonio Banderas, Salma Hayek, Joaquim de Almeida
Director	Robert Rodriguez
MPAA	R (Restricted)
Average Rating	4



Product details

Actors: Antonio Banderas, Salma Hayek, Joaquim de Almeida, Cheech Marin, Steve Buscemi

Directors: Robert Rodriguez

Writers: Robert Rodriguez

Producers: Carlos Gallardo, Robert Rodriguez, Bill Borden, Elizabeth Avellan

Format: PAL, Special Edition, Widescreen

Language English

Region: Region 2 (DVD formats.)

Number of discs: 1

Classification: 18

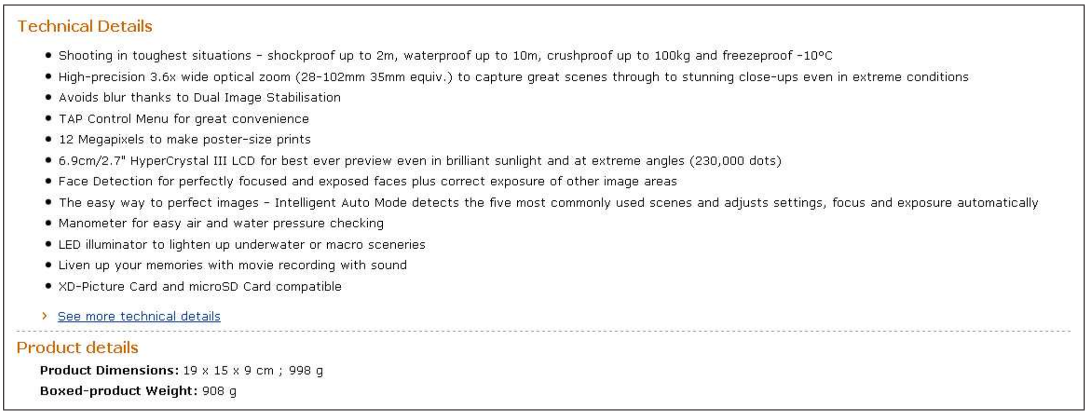
Figure C.2: Sample of data available on Amazon website for the movie “Desperado”, retrieved March 2009

shows an example of the types of features that are available via Amazon for cameras.

These two considerations were important for our experimental design (Section 6.6.5). The distribution of features was required so that there would be a variance in the ratings for cameras (just like we controlled for variance in ratings in the second movie experiment). The constraint on a minimum of three user reviews was important for our chosen baseline: the bar charts would be too uninformative for an interesting comparison with other explanations. In fact, a surprisingly large amount of cameras had zero reviews. While the cameras were manually inserted, the added features would have been available via the webservice as well.

C.2.8 Additional notes

This section is not meant as a comprehensive guide to Amazon services, but to give ideas as to what can be done with them. Although it may be more simple to do a basic search, it is not as clear which values are available in a response, or what type of response group to request. Likewise with Browsenodes, it is more difficult to grasp the relative position in



Technical Details

- Shooting in toughest situations - shockproof up to 2m, waterproof up to 10m, crushproof up to 100kg and freezeproof -10°C
- High-precision 3.6x wide optical zoom (28-102mm 35mm equiv.) to capture great scenes through to stunning close-ups even in extreme conditions
- Avoids blur thanks to Dual Image Stabilisation
- TAP Control Menu for great convenience
- 12 Megapixels to make poster-size prints
- 6.9cm/2.7" HyperCrystal III LCD for best ever preview even in brilliant sunlight and at extreme angles (230,000 dots)
- Face Detection for perfectly focused and exposed faces plus correct exposure of other image areas
- The easy way to perfect images - Intelligent Auto Mode detects the five most commonly used scenes and adjusts settings, focus and exposure automatically
- Manometer for easy air and water pressure checking
- LED illuminator to lighten up underwater or macro sceneries
- Liven up your memories with movie recording with sound
- XD-Picture Card and microSD Card compatible

› See more technical details

Product details

Product Dimensions: 19 x 15 x 9 cm ; 998 g

Boxed-product Weight: 908 g

Figure C.3: Sample of technical features for the camera “Olympus Mju”, retrieved March 2009

a structure. One useful reference is Amazon’s API documentation[7]. This documentation is more geared toward REST calls however, and is not the best starting point for SOAP requests.

C.3 Getting user preferences

This module was also implemented in Java, but using the Netbeans IDE which was deemed suitable for rapid web development. The front end visible to the user is a java server page. The module consists of first retrieving user preferences in order to construct a simple user model. User preferences are sent as form request data to a servlet and are stored in a java bean. The java bean represents the user model, and stores the importance a user places on domain specific features in ranked order, as well as any additional nominal (unranked) features.

User preferences are considered w.r.t. the properties of the items in the database in order to perform selection (Section C.4). Once an item is selected, an explanation is generated for it (Section C.5).

C.3.1 Movies

The user model for movies stores which genres the user prefers and dislikes. It also compares the importance users assign to the features: average rating, MPAA rating (e.g. rated PG), actors and director. It also stores names of favorite actors/actress and directors. The user model for movies makes sure there are no conflicts between liked and disliked genres, and splits genres into three categories: liked, disliked, and acceptable (neither liked nor disliked). Screenshots of how user preferences for movies are inputted can be found in Appendix B.5.

[7] Amazon.com/developers, retrieved 2009

For our running example let us presume that the user specifies their preferences as follows. First, the user specifies their genre preferences, and states that they feel like action and adventure (recorded as a single genre), and thriller movies, but do not want to be recommended comedy movies. Next, they rate the features average, director, actors, and MPAA, with the respective scores (on a five point Likert scale) 2, 3, 4, and 2. In the final screen they specify that Antonio Banderas is one of their favorite actors.

The user model then interprets the input. The genres action and adventure and thriller are recorded as preferred genres, comedy as disliked, and the remaining possible genres as neutral. The other features are ordered by the importance the user gave them (ties are resolved by which occurs first) actors (4), director (3), average rating (2), MPAA (2).

C.3.2 Cameras

The user model for cameras ranks the features: zoom, resolution, price, weight, brand and camera type (SLR or point and shoot). Screenshots of how user preferences for cameras are inputted can be found in Appendix B.6.2.

C.4 Selecting items

This module is responsible for selection of items. In theory, this would be where recommendations occur. However, in our implementation it is largely a placeholder. The most common usage is to retrieve all the items from the database, which are then returned as a randomized hashmap. Alternatively, constraints can be put on the retrieved items. The only such implemented usage is genre restriction for movies: only retrieving movies that are liked or acceptable to a user. This has been used in the final evaluation described in Chapter 7, to ensure that participants are not asked to watch a movie they may dislike. For our example this would mean that none of the recommendations could be comedies, as the user had specified they did not want to be recommended any movies in this genre.

We envision two ways in which a recommendation engine could be added to the existing architecture. Firstly, the feature preferences specified by the user could be used in a knowledge-based recommender system to guide the *selection* or *strength of recommendation* for items. A simplistic implementation would compute the strength of the recommendations as a weighed sum of the important features (for that user) that occur for the observed item. This approach would suffer from the weaknesses of knowledge-based systems such as not being dynamic (see also Section 2.1.3). It might also be difficult to find enough or any items that fit the requirements (e.g. for obscure actors), and would also be limited by the coverage of the features (e.g. there is no feature that covers the genre "costume dramas").

Secondly, the recommendation algorithm would compute recommendation strengths for each item, and return a list of ranked items. The explanation mechanism then explains each presented item according to the meta-data available to it. The explanations in this case would be aimed at decision support rather than providing transparency, as the explanations are decoupled from the recommendation algorithm. A variation on this alternative would be to return a list with less strict ranking, allowing the recommendations to result in less positive explanations (e.g. *"this movie belongs to one of your disliked genres..."*). A limitation of this approach is that the relevant data might be missing, e.g. the name of the director is unknown. Note that we have chosen not to address additional concerns such as diversity of recommendations, novelty and serendipity etc.

C.5 Generating the explanations

C.5.1 Natural language generation

Natural language generation, henceforth referenced to as NLG, is a subfield of artificial intelligence and computational linguistics and more specifically, Natural Language Processing (NLP). NLG systems produce texts that are understandable as well as appropriate in a language, commonly English, from non-linguistic input. Hence the term *natural* refers to the type of language used by humans, rather than formal and programming languages. One of the most common architectures used in NLG is the pipeline architecture, and has three components (Reiter and Dale, 2000; Reiter and Dale., 1997; Jurafsky and Martin, 2000):

- **Document Planning -** Document planning decides the content and structure of the generate text.
- **Microplanning/Discourse planning -** The microplanner decides how information and structure should be expressed linguistically, such as how long sentences should be.
- **Surface realization -** Generates the actual, and grammatically correct, text from the linguistic structures created in the two previous phases.

C.5.2 SimpleNLG

Part of the reason this testbed was implemented in Java, was ease of web-development: a web-based testbed allows more flexibility for testing. The other main reason was the availability of a Java library for realizing natural language: SimpleNLG version 3.4 [8].

[8] http://csd.abdn.ac.uk/˜ ereiter/simplenlg/, retrieved January 2009

This library is not a complete natural language generation system, in that it does not represent the full pipeline architecture. It only performs the final step of realization, and handles combinations of parts of a sentence, punctuation etc. It also manages simple syntactic requests such as tense (e.g. past, present, future) and negation.

This means that the developer needs to have a good idea of the meaning (semantics) and content of what they want to say. In terms of making proper sentences from data retrieved from Amazon this library is ideal. Values are inserted in pre-prepared sentences that can be combined together flexibly using SimpleNLG. This means that a large number of sentences and explanations can be generated on the fly. SimpleNLG also takes care of factors such as aggregation (e.g. whether to use "and" or just a comma between two clauses/sentences), punctuation and capitalization.

C.5.3 Movies

The explanations for movies in all three experiments are very similar is structure. The example explanations mentioned here are based on the second movie experiment unless otherwise stated. Screenshots of explanations for movies can be found in Figures 6.1, and 6.4.

Baseline

This explanation is simple and says whether a movie is popular. In this case a single variable can be modified to denote whether or not it is popular, or if it belongs to the top 250 in the IMDB. The output is a TextSpec that is set as a sentence which means that it can be combined with other sentences if need be. Using our example this would result in an explanation that simply states: *"This movie is not one of the top 250 in the Internet Movie Database (IMDB)."*.

Non-personalized explanations

Both feature based explanations for movies consist of two parts: the genre explanation and the top feature explanation. Initially the genre explanation was identical in the personalized and random-choice condition (see below). For the second experiment and final evaluation, the genres in the non-personalized condition were stated without any evaluative judgment (e.g. good, bad).

Let us return to our example. For this user the sentence describing the genre would mention all the genres for a movie, but not relate them to the user's preferences: "This movie belongs to the genre(s): Action & Adventure and Children". This part of the explanation initially only considered a single genre for a movie at a time. Later versions of the testbed considered multiple genres. Here is an example of how the genre explanation was constructed using simpleNLG:

```
// Genre explanation, non-personalized
if (cond == Constants.COND_NONPERSONALIZED){
  NPPhraseSpec movie = new NPPhraseSpec("movie");
  movie.setDeterminer("this");
  s1.setSubject(movie);
  s1.addHeadModifier("belongs to the genre(s):");
  List cur_genres = genres.getGenres();
  Iterator genreIt = cur_genres.iterator();
    while (genreIt.hasNext() && genreIt!=null){
        Genre g = (Genre)genreIt.next();
        s1.addComplement(g.getPrettyGenre());
     }
firstExplanation = new TextSpec(s1);
}
```

The sentence describing the second feature describes one feature of the movie, selected at random. We also ensure that the selected feature is different from the one feature most highly rated by the user. When the selected feature is director, only one name is mentioned, while for cast up to three leading actors are mentioned. In our example, this sentence would not describe the actors, but another feature such as average rating: “On average other users rated this movie 4/5.0.”

Personalized explanations

In the personalized condition the genre explanation states which genres are preferred, acceptable, or disliked (see also Section C.3). For the personalized explanations, the other feature mentioned was the one the user ranked the highest. For the feature actors, at most three actors were listed.

In addition, if the top feature was additionally tuned to the user’s preferences this was mentioned. Our example user found acting most important, and the movie stars their favorite actor Antonio Banderas, so this is mentioned explicitly: e.g. *“This movie stars your favorite actor(s): Antonio Banderas.”* If the actor(s)/actress(es) is not a favorite this modifies the explanation to e.g. *“This movie stars: “Selma Hayek”*. A similar explanation is given if the most highly ranked feature is director, and the movie is directed by one of their favorite directors: *”Tim Burton directed this movie.”*

C.5.4 Cameras

These explanations use a similar structure to the explanations for movies, but have naturally required a few modifications (see also Chapter 6.6). Screenshots of explanations for cameras are shown in Figure 6.7.

Baseline

The baseline in this domain is a barchart. We used Cewolf, a javascript library which can be used inside a Servlet/JSP based web application to embed complex graphical charts [9]. The output is a barchart with three bars, one with the number of good reviews, one for ok reviews and one for bad.

Non-personalized

The non-personalized explanations use a fixed set of three features: price, brand and type. This selection of features is based on a brief pilot study (see 6.6.2). Given the flexibility of simpleNLG, it was simple to put together sentences describing multiple features, and only slight modifications to the testbed were required:

```
/* makeSecondExplanation - explanation for condition 2 feature based, but not personalized */
public TextSpec makeExplanation2(int i,
    ArrayList topFeatures, long userID, int nRatings){

    secondExplanation = getPriceExplanation(i);
    secondExplanation = new TextSpec(secondExplanation, getBrandExplanation(i));
    secondExplanation = new TextSpec(secondExplanation, getCameraTypeExplanation(i)
    );
    return secondExplanation;
}
```

Personalized

The personalized explanation also mentions three features, but the three that are ranked the highest in the user model.

[9] http://cewolf.sourceforge.net/new/index.html, retrieved January 2009

C.5.5 Realizer

As previously mentioned, each sentence is constructed into the internal representation of a "TextSpec", and combined when necessary. Realization transforms this into text, which is then returned to the user, formatted in HTML and embedded into the Java server page. Examples of completed explanations can be see in Figures 6.1, 6.4 (for movies), and 6.7 (for cameras).

C.5.6 Data logging

All experimental data was logged using a MySQL database. Two tables were dedicated to each experiment, with one for user demographics, and the other for the actual experiment data. The demographics tables stored basic demographics such as age and gender, but also genre, actor, and director preferences, the top feature for that user, and the condition to which they were assigned. The tables for experiment data store all the movie and explanation ratings, but also qualitative comments, timestamps for each response, the explanation presented to the user, and the item title (for movies).