**Juha Leino**

# User Factors in Recommender Systems: Case Studies in e-Commerce, News Recommending, and e-Learning

**ACADEMIC DISSERTATION**
To be presented with the permission of the School of Information Sciences of the University of Tampere, for public discussion in the Pinni auditorium B1097 on August 29th, 2014, at 12:00.

School of Information Sciences
University of Tampere

Dissertations in Interactive Technology, Number 17
Tampere 2014

## ACADEMIC DISSERTATION IN INTERACTIVE TECHNOLOGY

| | |
|---|---|
| **Supervisor:** | Professor Kari-Jouko Räihä, Ph.D.<br>School of Information Sciences,<br>University of Tampere,<br>Finland |
| **Opponent:** | Assistant Professor Katrien Verbert, Ph.D.<br>WISE Research Group,<br>Vrije Universiteit Brussel,<br>Belgium |
| **Reviewers:** | Prof. Dr. Dietmar Jannach<br>Department of Computer Science,<br>Technische Universität Dortmund,<br>Germany<br><br>Assistant Professor Martin Schmettow, Ph.D.<br>Department of Cognitive Psychology and Ergonomics,<br>University of Twente,<br>The Netherlands |

Acta Electronica Universitatis Tamperensis; 1452
ISBN 978-951-44-9551-9
ISSN 1456-954X.
http://tampub.uta.fi.

**Dissertations in Interactive Technology, Number 17**

School of Information Sciences
FIN-33014 University of Tampere
FINLAND

ISBN 978-951-44-9443-7
ISSN 1795-9489

Juvenes Print – Suomen Yliopistopaino Oy
Tampere 2014

# Abstract

Recommender systems are widely seen as an effective means to combat information overload, as they help us both narrow down the number of items to consider (e.g. by suggesting books we are likely to find interesting) and decide which item(s) to choose (e.g. which book(s) to buy). In effect, they are seen as helping us make better decisions at lower transaction cost. Consequently, recommender systems have become omnipresent in e-commerce and are also increasingly used in services in various other domains both online and offline where the number of items exceeds our ability to consider them all individually.

The research on recommender systems has largely focused on evaluating them in system-centric terms while ignoring user-centric aspects. This approach is now seen as having been detrimental to the field, as recommenders that are excellent in system-centric terms are not necessarily the ones that actually serve their users the best. Today, it is widely acknowledged that, while algorithmic aspects are important, it is the user-centric factors that make-or-break a recommender system by determining its adoption and use.

The publications that constitute this dissertation focus on user-centric aspects of recommender systems in three domains. We look at how actual users find and choose items on Amazon, the pioneering e-commerce player (Publications I and II); what kinds of underlying dynamics are involved in news recommending (Publication III); and various issues involved in employing recommender systems in e-learning, using a system that we designed and implemented (Publications IV–VII).

All studies, except Publication III that is a survey study, involve actual users using actual systems in authentic use context. Also, in all studies except III, we collected both subjective data (interview or questionnaire data) and observation data (over-the-shoulder observation or use log data), allowing us to minimize the effects of the say-do problem. Consequently, the publications offer us windows into the actual use of recommender systems in the context of the wholes of the services (rather than focusing on isolated aspects in exclusion of the whole). While the explorative nature of the studies does not always allow us to generalize the results widely, the studies provide us with views of the actual use dynamics that have received little research attention thus far, due to the difficulties involved in studying them.

The results underline that there are few universal truths as far as user-related factors in recommender systems are concerned. In fact, what is true in one context can be the opposite in another context due to changing user tasks and goals. For example, while minimizing user effort in eliciting preferences is often used as a guideline for e-commerce (so as not to distract users from their current activity), in e-learning the same process represents a learning opportunity not to be wasted. In fact, increasing the cost and complexity of evaluating items in e-learning appears to enhance honesty, learning, and user experience. Also, as motivations behind dishonesty are different in e-commerce and e-learning, engendering positive behavior in various domains needs to be tackled differently. Moreover, recommender systems and the complex contexts where they are employed easily lead to emergent behavior that is hard to predict. Consequently, in developing recommender systems that truly serve their users, user-centric testing is a must.

In addition, while recommender systems have been subject to intense research interest, there are competing definitions of what the term *recommender system* actually comprises. We review existing definitions of the concept and discuss their various merits, as well as the consequences of adopting them. In effect, we argue that the technology-centered definition of recommender systems unnecessarily limits the recommender system research, while a wider definition allows us to consider all facets of recommender systems.

# Acknowledgements

I would like to gratefully acknowledge and thank Professor Kari-Jouko Räihä for his supervisory work. Also, I would like to thank my co-authors, Professor Kari-Jouko Räihä (Publications I–III), Ms. Sanna Finnberg (Publication III), and Mr. Tomi Heimonen, Ph.D. (Publication VII).

I would also like to gratefully acknowledge and thank the reviewers, Prof. Dr. Dietmar Jannach and Assistant Professor Martin Schmettow, Ph.D., for their insightful reviews. Likewise, I would like to gratefully acknowledge and thank my opponent, Assistant Professor Katrien Verbert, Ph.D.

Tampere, July 9, 2014

*Juha Leino*

# Contents

# List of Publications

This dissertation consists of a summary and the following original publications, reproduced here by permission.

# Author's Contribution to the Publications

The research that resulted in Publications I and II was designed and conducted by the author. Professor Räihä supervised the research design. The results were analyzed by the author. The first drafts of the publications were written by the author and then finalized together with Professor Räihä.

The research that resulted in Publication III was designed and conducted by the author and Ms. Finnberg. Professor Räihä supervised the research design. The results were analyzed by the author who also wrote the first draft of the paper that was then finalized together with Professor Räihä and Ms. Finnberg. Professor Räihä also contributed some statistical work for the final version of the paper.

The research work that resulted in Publications IV–VI was designed and conducted the author. The results were analyzed by the author who also wrote the papers.

In the case of Publication VII, a part of the same research effort that resulted in Publications IV–VI, Dr. Heimonen conducted the user interviews that we had designed together and did the first analysis of their results. Otherwise, the study was designed and conducted by the author who also undertook all other analytical work. After the author had written the first draft of the paper, it was finalized together with Dr. Heimonen.

# 1 Introduction

"As of the mid-1990s the lesson has still not been learned. An 'information superhighway' is proclaimed without any concern about the traffic jams it can produce or the parking spaces it will require. Nothing in the new technology increases the number of hours in the day or the capacities of human beings to absorb information. The real design problem is not to provide more information to people but to allocate the time they have available for receiving information so that they will get only the information that is most important and relevant to the decisions they will make. The task is not to design information-distributing systems but intelligent information-filtering systems."– Herbert Simon (1996, p. 144)

Recommender systems, or recommendation systems, recommendation agents, shopping agents, or shopping bots, as they have also been called interchangeably in research (Xiao & Benbasat, 2007), are widely seen as effective means to combat information overload (e.g. Lam, Frankowski, & Riedl, 2006; Farzan & Brusilovsky, 2006; Schafer et al., 2007; Xiao & Benbasat, 2007; Konstan & Riedl, 2012) in addition to, and partly because of, providing serendipity and discovery by guiding and complementing searching (Herlocker et al., 2004; Hangartner, 2007; Cremonesi, Garzotto, & Turrin, 2012a; Konstan & Riedl, 2012). Recommender systems help in two partially overlapping ways: First, by helping us find items that are likely to interest us (e.g. salient books on Amazon.com), typically by suggesting items; and second, by helping us decide which item(s) to choose (e.g. which book(s) to buy from the suggested books), often by providing information, e.g. community opinion and critiques (Schafer, Konstan, & Riedl, 2001). In effect, one of the primary uses of recommender systems is to assist us in making better decisions (e.g. Häubl & Trifts, 2000;

Schafer, Konstan, & Riedl, 2001; Lam, Frankowski, & Riedl, 2006; Hu, Liu, & Zhang, 2008; Castagnos, Jones, & Pu, 2009; Mudambi & Schuff, 2010).

In a sense, recommender systems seek to replace or augment the natural social process of asking friends, family, and knowledgeable persons or checking out, for example, movie or restaurant reviews in the newspaper to help us make decisions, and are, in that sense, modeled after informal word-of-mouth (Resnick & Varian, 1997; Schafer, Konstan, & Riedl, 2001; Swearingen & Sinha, 2002). Recommender systems can also be seen as having been modeled after a knowledgeable salesperson that assists consumers and thereby saves their time and effort (Chen, 2008).

Typically, a recommender system collects preference data from its users and then aggregates this information for its users (Resnick & Varian, 1997). In practice, the recommender systems research community has embraced both the systems that recommend items to users, often in the form of lists of suggested items that are personalized to the user, and the systems that help users evaluate items, therefore covering systems that provide *recommendations*, that provide *predictions*, and that provide *community opinion* (Schafer, Konstan, & Riedl, 2001). In effect, this definition covers a great variety of recommender systems, ranging from automated and personalized systems that "technologists" (Schafer, Konstan, & Riedl, 2001) often see as the best or, occasionally, as the only proper approach to recommender systems to non-personalized ones that are presented the same way to all the users of the system (these are manual in the sense that users need to read and evaluate user reviews to draw a conclusion about an item) (Schafer, Konstan, & Riedl, 2001; Lam, Frankowski, & Riedl, 2006).

Given the virtually infinite shelf-space available to e-retailers such as Amazon.com, they are in a position to offer consumers vast numbers of products from which to choose (Häubl & Trifts, 2000). At the same time, however, consumers have lost the ability to process all the possible items and alternatives, leading to a need of mass-customizing the presentation in e-stores (Häubl & Trifts, 2000; Schafer, Konstan, & Riedl, 2001; Burke, 2002). Jeff Bezos, CEO and founder of Amazon, has crystallized this drive for personalization by saying that if he had three million customers, he should have three million stores (according to Schafer, Konstan, & Riedl, 2001).

In fact, Resnick and Varian had already observed in 1997 that large numbers of e-retailers were introducing recommender systems for a wide range of products. Today, recommender systems are omnipresent in the larger online stores, playing an increasingly important role in e-commerce (Thor, Golovin, & Rahm, 2005; Said et al., 2012; Bobadilla et al., 2013; Jeckmans et al., 2013). Online businesses need to make their case in 10–20 seconds, and recommender systems are used to convince potential customers that it is worth their while to stay and take a look around the e-store, that the e-store offers value to them (Hangartner, 2007). The success of an e-commerce site can depend on its recommender systems, as the purchase decisions are increasingly made in online environments (Xiao & Benbasat, 2007; Castagnos, Jones, & Pu, 2009). In fact, today electronic word-of-mouth (eWOM), typically considered to consist of customer ratings and reviews, influences customer decisions more than any other media (Cui, Lui, & Guo, 2012; Racherla & Friske, 2012).

In e-commerce, recommender systems are considered to provide benefits to both consumers and e-retailers, i.e. the recommendation providers (Pu, Chen, & Hu, 2011; Cremonesi, Garzotto, & Turrin, 2012a). For consumers, recommenders provide means to reduce uncertainty inherently involved in purchase decisions, as consumers typically have to make purchase decisions with incomplete information (Hu, Liu, & Zhang, 2008; Mudambi & Schuff, 2010). In effect, recommender systems are to provide users with the necessary confidence to make decisions by reducing uncertainty (Hu, Liu, & Zhang, 2008; Castagnos, Jones, & Pu, 2009; Mudambi & Schuff, 2010; Cremonesi, Garzotto, & Turrin, 2012a). Recommenders also bring value to consumers by winnowing down the number of items to browse and reducing the decision time, therefore reducing decision costs (Pathak et al., 2010).

Still, these benefits do not come without considerable downsides. There are, for example, serious privacy and security issues connected to the relentless collection of data connected to user modeling that are often underestimated (Schafer, Konstan, & Riedl, 2001; Schafer et al., 2007; Konstan, 2008; Anderson & Rainie, 2012; Jeckmans et al., 2013). The dangers are underlined when such a serious player as Max Levchin, co-founder of PayPal, is building a machine that, according to his words, "knows more about you than you know about yourself" (O'Brien, 2006). However, in this dissertation, privacy and security issues are omitted from discussion because the publications do not deal directly with them.

For service providers, or e-retailers, recommenders provide increased revenue through their ability to persuade—to influence user attitudes, beliefs, decisions, and behavior (Gretzel & Fesenmaier, 2006; Castagnos, Jones, & Pu, 2009; Pu, Chen, & Hu, 2011; Cremonesi, Garzotto, & Turrin, 2012a) Recommenders can convert browsers into buyers, increase cross-selling, build loyalty, and diversify purchases (Schafer, Konstan, & Riedl, 2001; Pathak et al., 2010; Buder & Schwind, 2012; Cremonesi, Garzotto, & Turrin, 2012a).

Although the potential of recommender systems in e-commerce is often emphasized, in fact they have been employed in a large and diverse range of domains and contexts to help users make better decision, including, for example, music, movies, e-learning, news, health, jobs, and even divorce negotiations (Pommeranz et al., 2012; Bobadilla et al., 2013). However, different domains place differing requirements on recommenders, and transferring recommender systems and approaches from one domain to another typically requires adapting them to the new context (e.g. Stefani, Vassiliadis, & Xenos, 2006; Tang & McCalla, 2004a; Drachsler, Hummel, & Koper, 2009; Ghauth & Abdullah, 2010; Santos & Boticario, 2010; Buder & Schwind, 2012).

For example, while maximizing selling by recommending items of interest is important in e-commerce contexts, in e-learning the focus is on learning and, therefore, the recommendations should be guided by educational objectives rather than by user interests alone (Drachsler, Hummel, & Koper, 2009; Tang & McCalla, 2009; Granić & Adams, 2011; Santos & Boticario, 2010). Learners are, in fact, ready to read uninteresting study materials that are pedagogically grounded (Tang & McCalla, 2004a, 2009) while e-commerce customers would not be amused to be recommended items that do not interest them.

Evaluating recommender systems has long focused on system-centric aspects, such as using statistical measures to evaluate algorithms with various accuracy metrics (e.g. Herlocker et al., 2004; McNee, Riedl, & Konstan, 2006a; Cremonesi et al., 2011; Knijnenburg, Willemsen, & Kobsa, 2011; Buder & Schwind, 2012; Knijnenburg et al., 2012; Pu, Chen, & Hu, 2012). Although the importance of user-centric aspects and factors was already recognized early in the recommender research (e.g. Shardanand & Maes, 1995; Swearingen & Sinha, 2001), the inherent difficulties connected to user-centric evaluation of recommender systems (Kumar & Benbasat, 2006; Redpath et al., 2010; Cremonesi et al., 2011; Knijnenburg, Willemsen, & Kobsa, 2011; Buder & Schwind, 2012; Konstan & Riedl, 2012; Cremonesi, Garzotto, & Turrin, 2013) have resulted in limited research work in such user-centric factors as user experience (Cremonesi, Garzotto, & Turrin, 2012a) despite the fact that it is today widely accepted that they play a crucial role in e.g. recommender adoption and user perceptions of usefulness (Herlocker et al., 2004; McNee, Riedl, & Konstan, 2006a; Cremonesi et al., 2011; Knijnenburg et al., 2012; Pu, Chen, & Hu, 2012). At the same time, the rapid spread of recommender systems has resulted in a situation where being able to "characterize user experience and users' affective attitudes toward the [recommender] technology has become extremely important" (Pu, Chen, & Hu, 2011).

User-related factors can only be investigated with studies that involve real users (Pu, Chen, & Hu, 2012). Also, such studies need to take place in an actual use environment, because in complex online environments, independent activities of large numbers of users result in emergent behavior that is impossible to trace back to any individual user and that is difficult to study without actually running the system (Lerman, 2007; Buder & Schwind, 2012). In terms of Harrison and Dourish (1996), recommender systems and other social navigation tools transform a *space*, or opportunity, into a *place*, or understood reality, and thereby also transform the way people act within it (Harrison & Dourish, 1996; Chalmers et al., 2004; Svensson, Höök, & Cöster, 2005). Consequently, "measuring user experience requires developing a system, including both algorithms and user interface, and carrying out field studies with long-term users of the system—the only reliable way of measuring behavior in a natural context" (Konstan & Riedl, 2012).

As a consequence of this slanted research emphasis, relatively little is known about how users perceive recommendations and how they react to and act on the information that recommenders present to them (Buder & Schwind, 2012). For example, there are many open questions about how the selection of recommended items can be influenced (Buder & Schwind, 2012). Also, we need to better understand what drives consumer perceptions of eWOM, and how consumers deal with uncertainty and find information that they deem both useful and trustworthy (Racherla & Friske, 2012). Moreover, research is necessary for understanding how to achieve positive social outcomes with recommenders, how to employ them so that they result in such outcomes, and how to allow groups to determine what outcomes are desirable (Konstan & Riedl, 2012).

In addition, many human-centric factors are domain-dependent, as user tasks in different domains can be different, and consequently need to be studied in the current target domain. For example, our work shows that dishonesty is motivated differently in e-learning from e-commerce, and trying to engender positive behavior needs to be approached differently in the two domains. This makes establishing guidelines that apply to all recommenders in all use contexts very challenging. For example, while Pu, Chen, & Hu (2012) give a guideline that the easiness of the preference elicitation process should be a design goal, in e-learning giving explicit feedback should be seen as a learning opportunity and not as a burden (Buder & Schwind, 2012).

Also, while the emphasis in recommender research is often on efficiency improvements that recommenders can bring forth, in fact using recommenders should also be fun and pleasurable (Harper et al., 2005; McNee, Riedl, & Konstan, 2006a; Knijnenburg et al., 2012). Recommenders have the potential to enhance the user experience in an environment by bringing in sociality through social markers, resulting in social presence, a feeling of not being alone in a space (Svensson, Höök, & Cöster, 2005). In fact, merely having customer reviews on a site can result in perceptions of social presence and improve perceptions of usefulness (Kumar & Benbasat, 2006). Social presence, in turn, can engender a sense of community among regular clientele (Mudambi & Schuff, 2010).

The publications that constitute this compound dissertation focus on user factors in recommender systems in three domains, namely e-commerce (Publications I–II), news recommending (Publication III), and e-learning (Publications IV–VII). Except for publication III, all studies have investigated genuine users using actual services and have combined subjective responses (collected with interviews or questionnaires) with observation (observing use as it takes place or through use-log data) to reduce the impact of the say-do problem. Also, except for publication III, all studies investigate wholes rather than some parts of the overall environment, allowing us to see the totality of the use rather than artificially investigating only some aspect or strip of the use. Taken together, they afford windows into the actuality of recommender use and how genuine users see and experience them. Publication III, a questionnaire study, discusses how people find news items online and what role recommendations play in the equation, laying bare the underlying dynamics of news recommending activity.

Since recommender systems form a vast field of research and the publications here cover three domains that all have their own special characteristics, we need to take a look at many various aspects of recommender systems before we discuss the publications and the contributions they make in the field in further detail. First, we turn to defining recommender systems. Perhaps surprisingly, given all the research interest that recommender systems have aroused, defining them is not a straightforward task, and two camps can be distinguished. Some researchers prefer a narrow, technology-centric definition that leaves out many areas, such as customer reviews, that other researchers consider an integral part of the field of recommender system research. After defining what we mean by recommender systems, we take a brief look at the history of recommender systems to see how it all began. Then we discuss the benefits that recommender systems offer to their users and providers. After this, we look at the two types of recommenders, first various types of algorithmic recommenders and then eWOM. We then turn to considering preference elicitation and how such input is motivated and collected. This is followed by a discussion on evaluating recommender systems and, in particular, the importance of the user experience in evaluating recommenders. Finally, we look at the special characteristics of the three domains that we cover in the publications. After discussing the domains, we look at the publications in detail and discuss their contributions and relate our work to the literature.

# 2 Defining Recommender Systems

Given the research interest that recommender systems have stirred up, one would think that giving an unambiguous, clear-cut definition of what recommender systems are would be a straightforward task. However, this is not the case. The definitions vary along several dimensions, in particular whether a recommender system only produces *personalized* recommendations, whether recommender systems are inherently *algorithmic* (i.e. algorithm is an integral and necessary part of a recommender system), and whether recommender systems support only finding items of interest or also assist in making decisions about them (e.g. choosing which one to purchase). To add to the confusion, some researchers adjust their definitions in different research contexts.

## 2.1 Original Definition of Recommender Systems

Before looking at various definitions available in the literature today, let us first look at the original definition by Resnick and Varian (1997) who coined the term *recommender system* in the first place (Neumann, 2007). They saw the term recommender system as being more general than the phrase *collaborative filtering* that was coined by Goldberg et al. (1992), the developers of the first recommender system, and that had been widely adopted in the literature (Resnick & Varian, 1997). Resnick and Varian (1997) felt that collaborative filtering as a concept was somewhat misleading, as "recommenders [i.e. those users who input evaluations] may not explictly [sic] collaborate with recipients, who may be unknown to each other" and as "recommendations may suggest particularly interesting items, in addition to indicating those that should be filtered out.” In effect, today the term *collaborative filtering* is delegated to denoting an algorithmic approach to generating recommendations [1] (Schafer, Konstan, & Riedl, 2001).

Resnick and Varian (1997) saw recommender systems as a means to “assist and augment” the “natural social process” of using recommendations of others to make choices when we do not have sufficient personal knowledge or experience of the alternatives. In a typical recommender system, user “recommendations” (e.g. a rating or a mentioning of an item) are aggregated and directed “to appropriate recipients”. In some cases, the aggregation represents the primary transformation and, in others, the value comes from matching “the recommenders” with “those seeking recommendations.”

Resnick and Varian (1997) placed recommender systems in a technical design space that is defined by five dimensions: recommendation/ evaluation content (e.g. a rating or a mentioning of an item), explicit/ implicit entry, anonymity level of the recommendation, aggregation approach, and use of recommendations. Recommendation/evaluation content could “be anything from a single bit (recommended or not) to unstructured textual annotations.” (Today, the word *recommendation* would be more commonly used to mean a recommendation presented to the user rather than the input to the system from a user, as it is used here.) Recommendations could be entered explicitly or the system could gather them implicitly: Users could explicitly e.g. rate items (e.g. GroupLens) or the system could implicitly e.g. mine mentions of URLs from existing bookmark folders (e.g. Siteseer) or monitor reading times (e.g. GroupLens). Recommendations could be anonymous or tagged with the recommender’s identity or pseudonym (semi-anonymous). Resnick and Varian (1997) considered the 4th dimension, aggregating evaluations, particularly interesting for exploration. Finally, evaluations could be used in various ways, e.g. to filter out negative recommendations, to sort items according to numeric evaluations, or to display the evaluations together with the items.

To sum up the definition by Resnick and Varian (1997) along the three dimensions, 1) recommender systems can display personalized or non-personalized output (i.e. recommendations), as displaying textual annotations with the evaluated item is non-personal if the same set is displayed to all users (Schafer, Konstan, & Riedl, 2011); 2) recommender systems can be algorithmic or non-algorithmic, as presenting textual annotations as they were entered is non-algorithmic while employing collaborative filtering is clearly an algorithmic approach; and 3) since textual evaluations may be displayed together with the item (e.g. customer reviews on Amazon.com), recommender systems can both help users find

---

[1] See Section 5.2 for details on collaborative filtering.

items of interest from the multitude of items and help them decide between items.

## 2.2 NARROW, OR TECHNOLOGIST, DEFINITION

Many papers, including many recent papers, define recommender systems more narrowly than Resnick and Varian (1997), emphasizing the algorithmic aspect, personalization of output, and limiting recommender systems to producing lists of items that the system judges to be of interest to the user. A good example of one such definition is that of Knijnenburg et al. (2012) that states unequivocally that the algorithm is "[a]n essential part of any recommender system" and that the algorithm "provides personalized recommendations", i.e. "offer[s] each user a personalized subset of items, tailored to the user's preferences", typically as "an ordered list of recommendations."

Knijnenburg and colleagues are far from being alone. Many others have also emphasized the algorithmic nature of recommender systems, e.g. Cremonesi et al. (2008) state that recommender systems "use statistical and knowledge discovery techniques in order to recommend products to users", Park et al. (2012) state that recommender systems "use analytic technology to compute the probability that a user will purchase one of the products at each place, so that users will receive recommendations for the right products to purchase", and Pu, Chen, and Hu (2011) also define a recommender system as "a software technology that proactively suggests items of interest to users based on their objective behavior or their explicitly stated preferences."

Likewise, many see personalization as inherent to recommenders; e.g. Pommeranz et al. (2012) see recommender systems as "tools that provide personalized recommendations to people", Buder and Schwind (2012) state that "recommender systems are personalized, i.e. they suggest items that are adaptively tailored to the needs, interests, and preferences of a user", and Cremonesi et al. (2008) consider the aim of a recommender system to be "to predict which items a user will find interesting or useful."

In fact, Burke (2007) claims that the "criteria" of recommender systems being personalized in addition to providing interesting and useful recommendations is what separates them from information retrieval systems or search engines, as search engines return all the items matching the query, typically ranked by how well they match it. Then again, Burke (2007) also states that such techniques as "relevance feedback enable a search engine to refine its representation of the user's query, and represent a simple form of recommendation", so the distinction is perhaps not that clear. On the other hand, searching typically requires user involvement while recommender systems work "proactively" to suggest items of interest (Pu, Chen, & Hu, 2011) and select "potentially relevant items" "automatically" to help users overcome the information overload (Gena et al., 2011). Knijnenburg et al. (2012) view the matter similarly, stating that, before recommender systems, users had, at best, such tools as searches and filters, while recommender systems allow them to receive personalized recommendations based on their preferences.

In the same vein, many feel that the output of recommender systems is a list of items, e.g. Häubl and Trifts (2000) state that a recommendation system "generates a personalized list of recommended alternatives", Jeckmans et al. (2013) maintain that a "recommender system provides a set of items (e.g. content, solutions, or other users) that is most relevant to a particular user of the system", and Campochiaro et al. (2009) write that recommender systems provide "users with a short list of items which they are certain to enjoy."

Furthermore, many emphasize that recommender systems are, by default, online (Häubl & Trifts, 2000; Pommeranz et al., 2012). In contrast, Konstan and Riedl (2012) state that "today recommender systems are embedded in a wide range of commerce and content applications (both online and offline)".

In summary, according to the narrow definition, a recommender system uses user preferences, gathered implicitly or explicitly, to generate a personalized list of recommended items for a user. This definition emphasizes an important part of the reason why we have recommender systems in the first place: Faced with an increasing numbers of items, we need technological tools to help us to deal with the information overflow, to help us find items that are likely to be of interest to us.

## 2.3 WIDER DEFINITION

However, recommender systems have also been defined more widely, especially by researchers connected to the University of Minnesota and GroupLens Research. Schafer, Konstan, and Riedl (2001) state that the term *recommender system* "refers both to systems that specifically *recommend* lists of products and to those that help users evaluate products." They maintain that the recommender systems research community embraces systems that provide *recommendations*, systems that provide *predictions* and systems that provide *community opinion*. Consequently, rather than just suggesting items to users, recommender systems also help them to decide between items (Schafer, Konstan, & Riedl, 2001). Lam, Frankowski, and Riedl (2006) nicely capture the spirit of the wider definition by stating that "[o]ne of the primary uses for a recommender system is to help people make decisions."

Looking at this definition, some recommender systems are automated and personalized to each user while others are non-personalized and manual in the sense that users need to read and evaluate item reviews—the "unstructured textual annotations" that Resnick and Varian (1997) mention—to reach a conclusion (Lam, Frankowski, & Riedl, 2006). In fact, although "[t]echnologists often assume that the 'best' recommender application is … fully automatic and completely invisible", that is not always the case (Schafer, Konstan, & Riedl, 2001). Schafer, Konstan, and Riedl (2001) classify such recommender applications that provide "identical recommendations to each customer" as *non-personal*. Non-personal recommendations include "Top-sellers, editor choices, average ratings, and unfiltered customer comments" (Schafer, Konstan, & Riedl, 2001).

Interestingly, some authors who typically subscribe to the narrow definition, stating that recommendations are always personalized to the particular user, occasionally allow that not all recommendations are personalized. Cremonesi, Garzotto, and Turrin (2012a) state that distinguishing between systems that offer personalized and non-personalized recommendations is a "key taxonomic criterion" and that non-personalized recommender systems "select items without taking into consideration the user profile." As quoted above, however, Cremonesi et al. (2008) stated elsewhere that the aim of the recommender systems is "to predict which items a user will find interesting or useful." Likewise, Knijnenburg, Reijmer, and Willemsen (2011) talk about non-personalized recommenders that provide the same recommendations for each user and even find that "novices and maximizers seem to benefit more from a non-personalized recommender that just displays the most popular items." Nevertheless, in 2012, Knijnenburg et al. state that the recommender systems "offer each user a personalized subset of items, tailored to the user's preferences", as discussed above. It seems that when some researchers wish to contrast personalized recommendations against non-personalized, e.g. TopN[2] in both examples here, they accept that there are non-personalized recommender systems, but otherwise they do not allow for non-personalized approaches. It goes without saying that consistent defining of recommender systems would be beneficial to the field.

Given that such recommender systems as customer reviews are basically presented as entered and used manually (Lam, Frankowski, & Riedl, 2006), they are clearly non-algorithmic vis-à-vis automatic approaches using e.g. collaborative filtering algorithms to select items for a user. Average ratings and TopN approaches involve some computation but, again, when compared to such approaches as collaborative filtering, they are clearly less algorithmic. If we accept "unfiltered customer comments" as

[2] TopN recommendations are lists of top items based on e.g. sales or the number of views or downloads that are shown to all users in the same way (i.e. non-personalized).

recommender systems, as Schafer, Konstan, and Riedl (2001) suggest, then we inevitably have to also admit that some recommender systems are non-algorithmic. Whether we can do with a binary taxonomy—algorithmic or non-algorithmic—or wish to see algorithmic involvement in the generation of the recommendations as a continuum—customer comments as non-algorithmic, average ratings as slightly algorithmic, and prediction computing as heavily algorithmic—is a question for later work. Here the purpose is to establish that there are indeed non-algorithmic recommender systems.

In fact, by counting "unfiltered customer comments" (Schafer, Konstan, & Riedl, 2001) as a recommender system, we have already established that the output of recommender systems is not limited to generating lists of recommended items, be they personalized or not (e.g. TopN). Herlocker et al. (2004), in their oft-quoted paper, identified two end-user tasks that are typically discussed within the collaborative filtering literature, namely *annotation in context* and *find good items*. Finding good items is obviously a core task for many recommender systems, the output of which is typically a ranked list of suggested items (Herlocker et al., 2004). However, when it comes to annotating in context, the goal is to help users tell apart desired and undesired items (Herlocker et al., 2004). Originally, in Tapestry (Goldberg et al., 1992) and GroupLens (Resnick et al., 1994), predictions were applied to structured message databases, thus requiring the order and message context to be retained (Herlocker et al., 2004). Consequently, predictions of how much the user would possibly like the item needed to be used to annotate messages in context (Herlocker et al., 2004). If the typical annotation in the beginning was a calculated prediction, very soon, in e-commerce, items were annotated in the context of item lists (pages listing items, e.g. search results on Amazon.com) and item pages (a page that contains detailed information for an item, e.g. a page that gives all details available on a book) with average ratings, customer reviews, tags, and so on. Again, these recommendation systems were there to help users evaluate products (Schafer, Konstan, & Riedl, 2001), to help them choose items for closer inspection and to choose item(s) for purchasing (or some other action, e.g. viewing, listening, or borrowing, depending on the use and task context).

One of the most distinctive differences between the narrow definition and the wider definition is that the recommender systems, as described in the narrow definition, do not provide negative recommendations; they never list items that the user is most likely not to be interested in, saying, in effect, that "Here are some items that you will not like." The wider definition, by including ratings and reviews, can also provide negative recommendations; e.g. a book with an average rating of one-star that is based on fifty reviews on Amazon.com is quite obviously negatively recommended. Also, if the recommending task is to provide *annotation in* *context* (Herlocker et al., 2004), by calculating a prediction of how likely the user is to like each listed item rather than listing only items the user is likely to be interested in, we can obviously end up with negative recommendations, as the prediction may be that we are unlikely to be interested in a particular item. However, if the recommender systems, by definition, only provide lists of recommended items, negative recommendations are not logically possible.

In summary, the wider definition encompasses both the systems that recommend items, often as lists of items, and the systems that help us evaluate the items, thereby helping us make decisions (Schafer, Konstan, & Riedl, 2001; Lam, Frankowski, & Riedl, 2006). While in essence it has the spirit and contents of the definition by Resnick and Varian (1997), it puts it more clearly and uses current recommender system vocabulary to do so. According to it, recommender systems help us deal with information overload in two somewhat overlapping ways, by helping us find salient items (e.g. books we might be interested in) and by helping us to make decisions (e.g. which book to buy) (Schafer, Konstan, & Riedl, 2001; Leino & Räihä, 2007; Leino, 2012a). Therefore, recommender systems are not limited to generating personalized lists of recommended items, although this certainly constitutes an important recommendation task and many recommendation systems deal with this. Hangartner (2007) captured the heterogeneity of recommender systems nicely in stating that "the recommender idea is a formal concept that has as many different concrete examples as there are separate market applications."

The wider definition of recommender systems also begets us to consider the relationship between recommender systems and electronic word-of-mouth (eWOM), namely whether eWOM is conceptually separate from recommender systems or comes under recommender systems. Before considering this, however, let us take a look at the general process of selecting items when assisted by recommender systems, as its stages are related to the two definitions of recommender systems and also to whether or not eWOM comes under recommender systems.

## 2.4 THE PROCESS OF FINDING AND CHOOSING ITEMS

The infinite virtual shelf space allows today's online stores, such as Amazon.com, to make huge numbers of alternatives available for consumers (Häubl & Trifts, 2000). While a large number of alternatives is desirable from the consumer viewpoint, the limited cognitive resources of human information processing make carefully evaluating all alternatives challenging if not downright impossible (Häubl & Trifts, 2000; Xiao & Benbasat, 2007). In effect, while e-commerce is growing rapidly and purchasing decisions are increasingly made in online environments, the human cognitive capacity has stayed the same (Simon, 1996, p. 144), resulting in online e-commerce environments becoming complex decision-making environments (Häubl & Trifts, 2000; Xiao & Benbasat, 2007). The same applies for many other online environments where the number of items is way beyond the human cognitive capacity to process them one by one, be they libraries, online newspapers, or something else. While large numbers of alternatives are positive in one sense, they simultaneously make finding salient items more difficult due to information overload.

Consequently, consumers are typically thought to use a two-stage decision-making process in choosing items to purchase or consume in order to reduce the complexity involved. At the first stage, or the initial screening stage, a large number of available items are screened to determine which ones are worth an in-depth consideration. Items are not examined to any great depth at this stage; the goal is to identify a subset, or the consideration set. At the second stage, the items in the consideration set are compared and examined in-depth before committing to an alternative, i.e. choosing an item. (Häubl & Trifts, 2000; Xiao & Benbasat, 2007)

However, while described here as two discrete stages, in fact the transition between the stages is not sequential—complete one and then move to another—but iterative by nature (Chen, 2010). Consumers, for example, may have selected a set of items and, while looking at them in greater detail, may notice an interesting product on the *related recommendations* (such as *Customers who bought…* on Amazon.com) and, after a cursory look, may add it to the consideration set (Leino, 2008). Gradually, however, the consideration set size decreases until a final decision is reached (Chen, 2010).

Moreover, while the process is typically considered to consist of two stages, "the initial screening stage and the in-depth comparison stage" (Xiao & Benbasat, 2007), Chen (2010) suggests that, in fact, there are three stages: 1) screening alternatives to select items for in-depth evaluation, 2) examining item details and saving it on a wish list if near-satisfactory, and 3) comparing items on the wish list and making the final choice. For our purposes here, however, the two-stage process is a clear way to discuss the process, especially given that we consider it iterative by nature[3].

Häubl and Trifts (2000) see a recommendation system as a system that "generates a personalized list of recommended alternatives", and consequently see recommender systems as assisting consumers in the first stage by helping them generate the consideration list more efficiently. They do not see recommender systems as being useful at the second stage,

[3] At the same time, however, it is important to keep in mind that the actual picture is more complicated than this (see Section 10.2 for how the two-stage process fails to describe the actuality of user behavior in Amazon). In general, models based on sequential actions tend to oversimplify cognitive processes and can lead to overgeneralizations (Dinet, Chevalier, & Tricot, 2012).

and suggest that a comparison matrix is a more suitable decision aid to employ there.

In effect, this is the inevitable conclusion when the narrow definition of recommender systems is adopted. If recommendations need to be personalized, presented as a list, and generated algorithmically based on user preferences, recommender systems can be only of slight assistance at the second stage, e.g. by annotating in context as to how much the system predicts the consumer to like a particular item.

However, if we adopt the wider definition and regard such features as average ratings and customer reviews as recommender systems, it is easy to see why Xiao and Benbasat (2007) conclude that "typical RAs [recommending agents, i.e. systems] facilitate both the initial screening of available alternatives and the in-depth comparison of product alternatives with the consideration set, they can provide support to consumers in both stages of the decision-making process."

Consequently, how we choose to define recommender systems effects our conclusions about how they assist us and therefore how they should be designed to assist us. It seems that there are conflicting conclusions concerning recommender systems in the literature that arise simply from the authors using differing definitions. Obviously, the reader better be aware and read the current definition carefully. Regrettably, not all authors carefully define what they mean by recommender systems, apparently feeling that others share their definition by default. As seen here, this is not the case, and authors are encouraged to make their thinking transparent also when it comes to how they define recommender systems, at least until there is a commonly accepted definition.

## 2.5 ELECTRONIC WORD-OF-MOUTH

When lacking decisive first-hand knowledge, humans have successfully used the opinions, knowledge, and experiences of like-minded, similarly situated people to guide their choices in a natural social process (Hill et al., 1995; Resnick & Varian, 1997). Traditional word-of-mouth communication "was originally defined as an oral form of interpersonal noncommercial communication among acquaintances" (Cheung & Lee, 2012), and was limited largely to local social networks, spreading through personal contagions (Hu, Pavlou, & Zhang, 2006).

The advent of the Internet has radically increased the scale, scope, and speed of word-of-mouth communication (Hu, Pavlou, & Zhang, 2006; Cheung & Lee, 2012). In effect, electronic word-of-mouth (eWOM) effectively represents "a paradigm shift in word-of-mouth communication," giving it global reach, great speed of diffusion, persistence, and measurability while, at the same time, leaving consumers to judge the credibility of the source based on the cues available on the online system in question (Cheung & Lee, 2012; see also Hu, Pavlou, & Zhang, 2006). eWOM communication enables people to share their opinions on and experiences with goods, services, and other types of items asynchronously on review websites (e.g. Epinions.com), retail websites (e.g. Amazon.com), discussion fora, newsgroups, blogs, and social networking sites (e.g. Twitter and Facebook) (Hennig-Thurau et al., 2004, Cheung & Lee, 2012). In effect, eWOM has enabled socially and geographically dispersed strangers to share opinions and experiences related to products and services (Cheung & Lee, 2012)—to form virtual communities where people influence each other as if interacting (while not interacting directly), and thereby saving on communication costs, e.g. the time costs of developing relationships, costs to privacy, and costs of synchronous communications (Hill et al., 1995).

Information sharing, such as eWOM, is generally viewed as a public good phenomenon. A *public good* is a shared resource that is available to every member of a group, independent of whether or not they personally contribute to its creation and whose availability is not diminished by use. (Cheung & Lee, 2012)

Capturing succinctly all these aspects, Hennig-Thurau et al. (2004) define eWOM as "any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet." eWOM has very quickly become a major source of information for modern consumers (Hu, Pavlou, & Zhang, 2006; Racherla & Friske, 2012) and an important shaper of consumer behavior, with over 60% of consumers consulting eWOM before purchasing a new product or service and 80% of those planning to make an online purchase consulting online customer reviews before making the purchase decision (Cheung & Lee, 2012). In fact, eWOM—when positive—may allow retailers to charge higher prices (Pathak et al., 2010; Cheung & Lee, 2012). Consequently, in addition to research interest in the field of information systems, eWOM is also being studied in various business disciplines, such as marketing, consumer behavior, and economics (Cheung & Lee, 2012).

eWOM literature largely concentrates on reviews and ratings; e.g. Pathak et al. (2010) exemplifies the eWOM service as a retail web-site "where shoppers can submit and share their feedback about products through reviews and ratings systems". While such interfaces as Epinions.com, where consumers can view the profiles of the reviewers in addition to the ratings and reviews of products, mimics rather "accurately the social process of word-of-mouth recommendations" (Schafer et al., 2007), it behooves to ask if all eWOM communication takes place in those formats.

Several researchers have, in fact, extended eWOM to algorithmic recommender systems, too. Recommender systems employing collaborative filtering approach, and user-based collaborative filtering in particular, have especially been seen as capturing "how word-of-mouth recommendations sharing works" (Schafer et al., 2007; see also Shardanand & Maes, 1995) and as mimicking word-of-mouth recommendations, since they "use the opinions of like-minded people to generate recommendations" (Xiao & Benbasat, 2007). In fact, Schafer, Konstan, and Riedl already saw in 2001 recommender systems in general as being "modeled after informal 'word-of-mouth'", as they "responded directly to consumers, giving them independent advice". Schafer et al. (2007) see online services as being able to go "beyond simple word-of-mouth" by not only enabling us to determine "what a much larger community thinks of an item" but also by being able to provide us with "a truly personalized view of that item using the opinions most appropriate for a given user or group of users."

However, while eWOM extends to other types of recommender systems than reviews and ratings, we cannot really equate recommender systems and eWOM, as not all recommender systems fall under eWOM. For example, content filtering recommender systems generate recommendations based on the user's preferred item attributes instead of using the opinions of other people, as e.g. collaborative filtering does (Xiao & Benbasat, 2007; Schafer et al., 2007). Consequently, it would be hard to make a case for such recommender systems as being eWOM.

Pathak et al. (2010) discuss differences between different types of eWOM. According to them, while ratings and reviews typically come from consumers who have "heterogeneous shopping patterns," collaborative filtering tends to be based on preference data coming from customers who have "homogeneous shopping patterns." Also, they state that recommendation type of eWOM is more useful for reducing search costs whereas ratings and reviews assist in choosing items. In other words, recommender type of eWOM is useful in the first stage of the decision-making process while ratings and reviews are more useful in the second stage. They conclude that recommender systems are "a valuable addition to the general digital word of mouth."

Overall, then, we conclude that eWOM can be seen as a part of the recommender research field[4]. eWOM clearly helps users evaluate items and provide community opinion, and are thus within the scope of

---

[4] However, we recognize that the equation can be seen the other way around, too; Pathak et al. (2010) state that "online recommender systems" constitute "a valuable addition to the general digital word of mouth." In effect, they see online recommender systems as being part of eWOM and therefore eWOM as the larger concept (Pathak et al. 2010). While we look at the equation from the viewpoint of recommender systems, we are, by and large, comfortable with both viewpoints.

recommender systems, as defined by Schafer, Konstan, and Riedl (2001). In fact, given that Resnick and Varian (1997) already stated when coining the term *recommender system* that recommender systems include "unstructured textual annotations" that are aggregated, it is easy to claim that eWOM has been part of the field from the very start.

## 2.6 RECOMMENDER SYSTEMS AND TAGGING

As we will not specifically discuss tagging and recommender systems elsewhere in this dissertation, we take the opportunity here to glance at tagging in the context of recommender systems before discussing its relationship to recommender systems, i.e. whether tagging can be seen as constituting a recommender system. Tags, i.e. user-contributed free-form, non-hierarchical labels that function as metadata, have traditionally been used in information systems in general for (social) navigation; remembering, finding, and sharing items and information; structuring and organizing items and information; and serendipitous browsing (Sen et al., 2006; Jäschke et al., 2007; Lee, Son, & Han, 2007; Gedikli & Jannach, 2010). Collaborative tagging systems are in use in numerous services (Gedikli & Jannach, 2010), e.g. Last.fm, Del.icio.us, Flickr.com, and Amazon.com. Tags and folksonomies (rich information spaces consisting essentially of sets of triples: user, item, and tag) that emerge from tagging have been used in recommender systems in various ways and in various roles, as have recommender systems been used to enhance tagging systems, too (Gedikli & Jannach, 2010; Bobadilla et al., 2013).

In effect, recommender systems have been used to recommend tags and tags have been used to enhance recommendation (Bobadilla et al., 2013). *Tag recommendation systems* help users find good tags and they have already been used since the early days of tagging systems; Del.icio.us, founded in 2003, for example, had already incorporated a tag recommender in 2005 (Jäschke et al., 2007). While tag recommenders assist users in the tagging process (coming up with good tags requires careful thought), they are also used to increase the chance of getting an item annotated, and, importantly, to consolidate the tagging vocabulary (Jäschke et al., 2007; Marinho & Schmidt-Thieme, 2007; Konstan & Reidl, 2012). Konstan and Reidl (2012) give as example three tags given on MovieLens to apparently denote the same director, *Tarantino*, *Quentin Tarantino*, and *tarantino is god*; having a wide number of nearly synonymous tags is not beneficial to the system, and recommender systems can help a tagging folksonomy converge (Sen et al., 2006; Konstan & Reidl, 2012).

Importantly, users with similar tagging vocabulary tend to tag alike, which implies and underlines a need to personalize tag recommendations (Marinho & Schmidt-Thieme, 2007). For example, Landia and Anand (2009) used document similarity in combination with user similarity to recommend personalized tags, managing to increase the precision score of the system.

Tags have also widely been used in the recommendation process, typically as additional knowledge (Tso-Sutter, Marinho, & Schmidt-Thieme, 2008; Gedikli & Jannach, 2010; Bobadilla et al., 2013). For example, tagging data has been used to identify similar users or viewed as additional information about the items to be recommended (Gedikli & Jannach, 2010). Tags have been used to enhance both content-based recommender systems (e.g. using tags as item descriptions) and collaborative filtering recommender systems (e.g. using tags to find similar users) in addition to hybrid systems (Gedikli & Jannach, 2010; Bobadilla et al., 2013). By exploiting "existing interaction between users, items and tags", such approaches have managed to improve algorithmic effectiveness both in terms of predictive accuracy and coverage (Gedikli & Jannach, 2010).

While tags naturally constitute additional information, or metadata, that can be used in content-based recommender systems, Tso-Sutter, Marinho, and Schmidt-Thieme (2008) have proposed a generic method that allows incorporating tags to standard collaborative filtering algorithms. They reduced the three-dimensional correlations (user-item-tag) to three two-dimensional correlations and then fused them to re-associate correlations.

Taking an innovative approach, Gedikli and Jannach (2010) integrated tags in recommender systems by having users rate items by rating the corresponding tags. In this way, they extended the work of Sen, Vig, and Riedl (2009) who had automatically inferred the user's preferences for individual tags—using MovieLens as the study environment, they thus determined to what extent the user liked the movies tagged with a tag to infer the user's preference for the tag. The rating prediction for a movie was then based on the aggregation of the inferred user preferences for the tags assigned to the movie (Gedikli & Jannach, 2010). They managed to generate more precise recommendations with their approach in comparison to several other algorithms and preference inference metrics (Gedikli & Jannach, 2010).

In effect, Sen, Vig, and Riedl (2009) had inferred global preferences for the tags—the user either liked movies annotated with the tag or not. Gedikli and Jannach (2010), in contrast, felt that tags were meaningful in the context of an item—in case of movies, for example, the user might like Arnold Schwarzenegger in *action movies* and not like him in *comedies*. Consequently, Gedikli and Jannach (2010) had users rate tags in the context of the item to further improve the accuracy. In this way, they also employed a multi-criteria or multi-dimensional approach, as users could use as many dimensions/criteria (tags) as they wanted and rate the item by them (i.e. rating items by rating the corresponding tags). However, allowing users to determine the rating dimensions with tags, while empowering them to have a more rich expression of their preferences, naturally results in less homogenous and less structured annotations (Gedikli & Jannach, 2010; Konstan & Riedl, 2012).

In effect, tags have been used to enrich recommender systems in recent years in various ways (Gedikli & Jannach, 2010) and the trend seems to continue to produce improvements.

#### Explaining Recommendations with Tags

Tags have also been used in an explanatory role in recommender systems. It appears that very simple explanations work the best (Herlocker, Konstan, & Riedl, 2000), and tags, in fact, offer an interesting and effective way of providing explanations (Vig, Sen, & Riedl, 2009; Konstan & Riedl, 2012).

Vig, Sen, and Riedl (2009) consider tags to have many qualities that make them effective for explanations: tags can "describe what an item is, what it is about, or what its characteristics are" by providing "both factual and subjective descriptions." They studied two aspects of tag-based explanations, *tag relevance* ("the relationship of the tag to the recommended item", or how well the given tag describes the item) and *tag preference* ("the relationship of the user to the tag", or the user's sentiment towards a certain tag). They saw tag relevance as item-specific and tag preference as user-specific.

In tagsplanation, the interface that Vig, Sen, and Riedl (2009) presented in their paper, a tag or a set of tags functioned as the intermediary entity; e.g. "*We recommend the movie* Fargo *because it is tagged with* quirky *and you have enjoyed other movies tagged with* quirky." Two types of features have been used as an intermediary entity in recommendation explanations in feature-based explanations, predefined categories and keywords (Vig, Sen, & Riedl, 2009). *Predefined categories*, while having a great potential, also have challenges, as experts are needed to create the categorization scheme, the scheme might not be expressive enough, and the categories may be too static to accommodate changing needs (Vig, Sen, & Riedl, 2009). In building a recommendation system that relies on categories, the *New York Times* found that they were spending a lot of "time and effort tagging contents" with "semantic tags that tell the recommendation engine what each story is about" (Krakovsky, 2011). On the other hand, *keywords* can be mined from textual contents and they can be used for explanations (Vig, Sen, & Riedl, 2009). Both approaches have been employed successfully to help users make more accurate decisions (Vig, Sen, & Riedl, 2009).

In contrast, tagging puts the user in control and allows free-form metadata—no expert is necessary to come up with the tags or to apply them to items (Vig, Sen, & Riedl, 2009). In terms of Tso-Sutter, Marinho, and Schmidt-Thieme (2008), categories are *global* descriptions of items while tags are *local* descriptions.

Vig, Sen, and Riedl (2009) found that both tag relevance and tag preference play a role in tagsplanations: tag relevance serves best in an organizing role, while tag relevance and tag preference both help accomplish the goals of *justification* (the ability of the explanation to help the user understand why an item was recommended), *effectiveness* (the ability of the explanation to help the user make good decisions), and *mood compatibility* (the ability of the explanation to convey if the item matches the user's mood). Overall, the study showed the viability of tagsplanations.

Gedikli, Ge, and Jannach (2011) studied using tag clouds as explanation interfaces in recommender systems. In a user study, they set a personalized tag cloud (tag preferences were shown by making the tags towards which the user had positive feelings blue and the ones towards which they had negative feelings red; neutral tags were in black), a non-personalized tag cloud (all tags in black, i.e. neutral), and a keyword-style explanation against each other. They found that tag cloud interfaces were perceived positively by users and that they could be used to make the explanation process more efficient and effective. In effect, tag cloud explanation interfaces helped users make better and faster decisions in comparison to keyword-style explanation interfaces. Considering that users favored even the non-personalized tag cloud interface over the keyword-style interface implies that other factors, possibly the graphical presentation that naturally affects the effectiveness of an explanation interface, played a role in the equation. However, personalized tag cloud interfaces were preferred over non-personalized tag cloud interfaces, indicating that personalization has an effect, too.

#### Tags as Recommendations

While tags clearly are preference data (Jeckmans et al., 2013) and have been, as discussed above, used to enhance recommendations in many ways and to provide explanations for recommendations, the question we now need to ask is if tags can be seen as a recommender system or not.

In fact, the answer seems to be both yes and no. Tags can be recommendations but not all tags are necessarily recommendations. Tags can be used to rate items and to express opinions (Lee, Son, & Han 2007) but they have many other functions, too. Many researchers have commented on the ability of tags to rate or review items (Lee, Son, & Han, 2007). For example, Marlow et al. (2006) listed *opinion expression* as one motivation that shapes tagging behavior, while Wal (2007) considered *rating* as something that tagging services offer, giving as an example the kinds of tags K-Fed's CD *Play with Fire* had garnered on Amazon.com—regrettably, Amazon.com has now removed tags, and so such tags as *AKA* *vogon poetry, awful*, and *hack*, not to mention a few less complimentary ones, are no longer there to be enjoyed.

Golder and Huberman (2006), based on Del.icio.us data, classified tags into seven categories according to the function they perform: Identifying what (or who) it is about, Identifying what it is, Identifying who owns it, Refining categories, Identifying qualities or characteristics, Self reference, and Task organizing. The category *Identifying qualities or characteristics* that includes adjectives such as *scary*, *funny*, *stupid*, and *inspirational* is used to "tag according to the tagger's opinion of the content" (Golder & Huberman, 2006). Such adjectives clearly rate—or mini-review (Leino, 2012b)—items. However, not all rating or reviewing tags need to be adjectives, as the tag *hack* for K-Fed illustrates.

Sen et al. (2006) collapsed the seven categories by Golder and Huberman (2006) into three more general classes that they related to user tasks that tagging could support on MovieLens: Factual tags, Subjective tags, and Personal tags. The most pertinent class to the current discussion is *subjective tags* that are used to "express user opinions" (Sen et al., 2006) that again brings us into the realm of rating and reviewing items.

It has been proposed that tags are more suited to expressing facts rather than opinions because they lack richness and structure for expressing complex opinion, thus falling somewhere between ratings and full-fledged comments (Cena et al., 2011). Still, as discussed above, tags are clearly used to express opinions on items, even if simple opinions, e.g. tag *hack* for K-Fed's deplorable CD. Also, if we accept that tags can express opinions and function as kind of mini-reviews, then tags can also be seen as part of the eWOM phenomenon, at least as far as their opinion-expressing function is concerned.

Using tags for ratings and opinion-expression, while clearly taking place, involves potential challenges (Lee, Son, & Han, 2007). Golder and Huberman (2006), for example, point out that tags used for rating and opinion-expressing tend to have diversified vocabulary, rendering them less useful for using them as metadata for the community-at-large. After all, tagging vocabulary is prey to various linguistic problems, such as polysemy, synonymy, and basic level variation (Golder & Huberman, 2006). In addition, Lee, Son, and Han (2007) point out that because tags are used for rating and opinion-expressing, there is a risk of redundant task imposition—many systems mingle tagging and rating, making users do the same task twice in a sense. On the other hand, as users often want to know the rationale behind the (numeric) rating (Leino, 2013), tags can help there as explanations, as discussed above.

Overall, as long as we go by the wider definition of recommender systems, we contend that a tagging system can be seen as a recommender system. If rating is a recommender system and tagging can be seen as being redundant with rating (i.e. the two can be used for the same task) (Lee, Son, & Han, 2007), then there is little space for argument on the matter. However, tagging has also many other functions e.g. related to information filtering and retrieval (Lee, Son, & Han, 2007) that do not really have a recommending function the way rating and mini-reviewing use of tags does. If tags are predefined and used to categorize items, they are not used as a recommender system (even if they naturally can be useful in finding items and can be used to direct recommendations to certain categories). In contrast, if users are asked to evaluate items with tags, then the tagging system is used as a recommender system. However, in many cases, both happen, as many collaborative tagging systems leave the tagging vocabulary up to the community. If we wish to encourage using tags for rating items, tagging offers potential for it (Lee, Son, & Han, 2007) but we probably need to instruct users to do so in order to make it a major function of the system. Consequently, we contend that some functions of tagging come under recommender systems and some do not. In other words, tagging research comes partially under recommender system research, but not entirely.

## 2.7 RECOMMENDER SYSTEMS AS INFORMATION SYSTEMS

Recommender systems fall within the domain of information systems (Xiao & Benbasat, 2007). Furthermore, recommender systems are decision-support systems in that they are information systems, they are used in decision-making, and they are meant to support humans, not to replace them (Svensson, Höök, & Cöster, 2005; Xiao & Benbasat, 2007).

At the same time, however, Xiao and Benbasat (2007) point out that recommender systems are different from traditional decision-support systems based on several aspects. Traditional decision-support systems were aimed at managers or analysts that used such systems for assistance in various tasks, typically in planning tasks, and, consequently, such systems typically employed process models. The users of recommender systems, on the other hand, come from all walks of life and typically face a class of problems known as *preferential choice problems*, and, consequently, recommender systems employ choice models and "support the integration of decision criteria across alternatives." (Xiao & Benbasat, 2007)

Recommender systems also share similar characteristics with knowledge-based systems, as they need to explain their reasoning to the users in order to engender trust in themselves and their recommendations (Herlocker, Konstan, & Riedl, 2000; Xiao & Benbasat, 2007). There is, after all, a kind of agency relationship between the recommender system and its user: The user (principal) cannot be sure if the recommender system is working solely for their benefit or, partially or entirely, for the merchant or the manufacturer that has made it available to use (Xiao & Benbasat, 2007). Users may have doubts about the benevolence of the systems in addition to having doubts about their abilities, and, consequently, engendering trust on the user is a key factor in recommender system adoption (Xiao & Benbasat, 2007; Knijnenburg, Reijmer, & Willemsen, 2011).

Importantly, Burke (2002, 2007) points out that recommender systems are not information retrieval systems. When a recommender system returns recommendations, they are understood to be items that are worthy of consideration, i.e. likely to be interesting or useful (Burke, 2002, 2007). In contrast, the items an information retrieval system returns are understood to be matches to the query—a search engine returns all matching items, ranking them by the degree of the match (Burke, 2007). Hangartner (2007) encapsulates the difference between search engines and recommender systems by stating that while search engines help us find what we know we are looking for, serendipitous discovery that recommender systems offer helps us find the rest. In a sense, a recommender system functions as an information agent for the user, often providing individualized recommendations that are relevant and interesting to the user (Gretzel & Fesenmaier, 2006; Burke, 2002). Still, while making a distinction between recommender systems and information retrieval systems is an important one to make, such search engines as Google do blur this distinction to some degree by incorporating authoritativeness to return more useful results (Burke, 2002).

## 2.8 Collective Intelligence: Wisdom of the Crowds

Recommender systems exploit the wisdom of the crowds by harnessing the crowds, giving birth to collective intelligence and self-organizing dynamics (Chen, 2008; Gedikli & Jannach, 2010; Malone, Laubacher, & Dellarocas, 2010; Buder & Schwind, 2012). *Collective intelligence* has been defined broadly as "*groups of individuals doing things collectively that seem intelligent*" by Malone, Laubacher, and Dellarocas (2010) from MIT's Center for Collective Intelligence. While collective intelligence has existed for basically as long as social groups have existed, be they humans, ants, or bacteria, the advent of the Internet has given rise to many new forms of collective intelligence (Malone, Laubacher, & Dellarocas, 2010). Moreover, the Internet has vastly magnified the power of the crowds (Chen, 2008).

Malone, Laubacher, and Dellarocas (2010) define answers to the questions *who*, *why*, *what*, and *how* as the building blocks, or genes, of collective intelligence. There are two possible answers to the question of *who* undertakes the activity, *hierarchy* and *crowd*. Most recommender systems are examples of *crowd*, as activities can be undertaken by "*anyone in a large group who chooses to do so*". In fact, much of the collective intelligence that is enabled by the Internet is based on *crowd* gene. Still, an editor's review could be seen as a recommendation that come from the *hierarchy* gene ("*someone in authority assigns a particular person or persons to perform the task*"), so it cannot be said that *all* recommender systems are based on the *crowd* gene. (Malone, Laubacher, & Dellarocas, 2010)

The question *why*—why people take part in the activity and what motivates and incentivizes them—is naturally closely related to the question *who* (Malone, Laubacher, & Dellarocas, 2010). While motivations to contribute will be discussed later in this dissertation in some detail, here we note that Malone, Laubacher, and Dellarocas (2010) give three basic genes: *Money* (direct payments or likelihood of future payments), *love* (intrinsic motivations, opportunities to socialize, or a feeling of contributing to a cause), and *glory* (recognition).

There are two genes to reply to the question of *what* is being done, *create* and *decide*. The two genes for the *decide* task, *group decision* and *individual decision*, are especially pertinent to recommender systems. In the case of *group decision*, inputs from a crowd "are assembled to generate a decision that holds for the group as a whole." In many recommender cases, this means "aggregating individual inputs to form a publicly visible estimate of a quality." Group decision includes *Voting* that can be explicit, e.g. thumbs-up or down on Digg.com, or implicit, e.g. *Frequently Bought Together* on Amazon.com. In the case of *individual decisions*, crowd members make decisions that are influenced by crowd input but are not necessarily identical. (Malone, Laubacher, & Dellarocas, 2010)

Collective decision making can be very effectual (Lerman, 2007). For example, the news of Rumsfeld's resignation in 2006 made it to the front page of Digg.com within three minutes of its submission, beating the algorithm of Google News by twenty minutes (Lerman, 2007). Another example is how the TV studio audience of *Who Wants to Be a Millionaire* found the correct answer 91% of the time while experts only found the correct answer 65% of the time (Chen, 2008).

One result of harnessing collective intelligence is that recommender systems exhibit emergent behavior, or emergent properties of groups, emerging from the decentralized actions and decisions of independent users (Lerman, 2007; Buder & Schwind, 2012); emergent patterns reflect the aggregate behavior of a crowd but the interaction of individual users are inherently *local* (Fu et al., 2010). In effect, recommendations (in the case of e.g. collaborative filtering) cannot be meaningfully traced back to the behavior of any one user, giving rise to group cognition (Buder & Schwind, 2012). Also, studies of tagging systems have found that "tagging distributions tend to stabilize into power law distributions", leading to communally discovered collective categorization schemes (Halpin, Robu, & Shepherd, 2007). As patterns emerge, community opinion becomes evident.

In fact, the use of collaborative intelligence transforms "both the space and the ways that people act within it" (Chalmers et al., 2004), in effect turning *space* into *place* in the terminology of Harrison and Dourish (1996) who define *space* as opportunity and *place* as understood reality. Halpin, Robu, and Shepherd (2007) suggest "treating collaborative tagging systems as complex systems" as this "yields important insights into the dynamics and processes of these systems." We suggest that viewing recommender systems the same way can reveal interesting emergent behavioral patterns that have not received sufficient attention thus far. Moreover, as Fu et al. (2010) underline, we need to understand the local interactions at the individual level to be able to "explain aggregate patterns at the social level." The aggregate, or social, level patterns influence the local level behavior from which aggregate patterns emerge, and we need studies of both the levels.

### Social Navigation

Recommender systems also represent a type of social navigation (Dourish & Chalmers, 1994; Dieberger et al., 2000; Chalmers et al., 2004; Svensson, Höök, & Cöster, 2005; Schafer et al., 2007). The term *social navigation* was coined by Dourish and Chalmers (1994) who saw it as "movement from one item to another" that is "provoked as an artefact of the activity of another or a group of others." Already in this early paper, Dourish and Chalmers (1994) stated that recommendations on people's home pages to other www-pages and collaborative filtering in systems such as Tapestry are part of social navigation. Social navigation allows people to "benefit from the presence and activity of others" (Ling et al., 2005) and Dourish and Chalmers (1994) saw collaborative filtering as "a means to exploit social knowledge to navigate around items of interest."

Prior to Dourish and Chalmers introducing the concept of social navigation, Hill et al. (1992) had introduced the concept of wear due to reading and editing, and had created the first history-enriched environment, a concept that is very much like the original concept of social navigation. However, the idea of sharing "trails" that users create through information spaces actually comes from Vannevar Bush's famous article *As We May Think* (Bush, 1945; Dieberger et al., 2000), a reworking of an earlier essay called *Mechanization and the Record* (1939)[5].

The concept of social navigation is based on the observation that watching and interacting with other people is part of everyday information seeking behavior (Svensson, Höök, & Cöster, 2005). Similarly, the idea that people tend to use social interaction with other people as a means of reducing uncertainty when selecting items was behind the concept of recommender systems; e.g. people often ask other people for recommendations when choosing a movie to see (Resnick & Varian, 1997). Both are largely about

[5] http://en.wikipedia.org/wiki/As_We_May_Think

people in a virtual community influencing each other as if interacting without actually interacting (Hill et al., 1995), although social navigation does also include more direct means of assisting in navigation, e.g. recommendations of links by email (Dieberger, 1997). As discussed above, such influencing can reduce communication costs in comparison to actual interacting (Hill et al., 1995).

Social navigation systems "exploit social practices and behaviour to help users navigate and explore" (Chalmers et al., 2004) by making the collective, aggregate, or individual behavior visible and useful as a basis for making decisions (Svensson, Höök, & Cöster, 2005; Schafer et al., 2007). The actions of others become visible in information traces that form social texture (Chalmers et al., 2004; Svensson, Höök, & Cöster, 2005; Lerman, 2007). Showing information about the choices that others have made affects both individual and collective decision making (Lerman, 2007; Chen, 2008).

Such social texture is dynamic by nature, as practices and choices change over time just like paths overgrow in forests if they are not used (Chalmers et al., 2004). For example, the books we see in *Customers who…* recommendations on Amazon.com change as the actions of other customers evolve. In this sense, social navigation is based on the way that spaces are transformed when people leave their marks on them: Social texture transforms a *space* into a *place* and affects how people act in it (Harrison & Dourish, 1996; Chalmers et al., 2004).

Social texture and the social cues that users leave behind them also help in turning the *place* social and pleasurable by engendering *social presence*, a sense of not being alone in the virtual space (Chalmers et al., 2004; Svensson, Höök, & Cöster, 2005). In addition to the social texture—including ratings, reviews, tags, and more algorithmic recommender systems—being useful in finding and selecting salient items and making the *place* more alive and inviting, it can also provide *social affordance*, awareness of what is appropriate behavior and what is not (Dieberger et al., 2000). The social cues left behind can become transformed into social practices, (virtual) rules and regulations, or even artifacts, e.g. signs and landmarks in the virtual *place* (Chalmers et al., 2004). Sociality and related aspects represent another way that social navigation, including recommender systems, can transform the environment and behavior in it—not only by assisting us in finding items of interest and making decisions but also by engendering social rules that affect behavior (Chalmers et al., 2004). For example, awareness of oversight on MovieLens has been shown to have a positive effect on the amount and quality of contributions (Cosley et al., 2005).

In effect, social navigation is not simply about helping users navigate more efficiently, or helping them find items of interest quickly, but through social presence it is also about making users stay a bit longer in the *place*, feel relaxed, and inspired to try things out, perhaps even purchase a product or service that they would not have otherwise considered (Dieberger et al., 2000).

Practically all that has been said here about social navigation applies to recommender systems. In fact, recommender systems are the prime exhibit of social navigation that we have today. Besides helping users find items of interest based on the behavior and preferences of the user community and providing community opinion to assist in decision-making, recommender systems also have an important role in providing social texture. Even many recommender system concepts and their labels directly underline the presence of others in the *place*; e.g. Amazon.com's *Customers who…* recommendation feature directly refers to the existence of other customers in addition to providing transparency about why the items are recommended.

## 2.9 SUMMARY

Recommender systems are information systems, and, more specifically, decision-support systems (Svensson, Höök, & Cöster, 2005; Xian & Benbasat, 2007). Furthermore, recommender systems harness collective intelligence and enable social navigation, thus also providing social presence and helping transform *spaces* into *places* (Dourish & Chalmers, 1994; Harrison & Dourish, 1996; Dieberger et al., 2000; Chalmers et al., 2004; Svensson, Höök, & Cöster, 2005; Schafer et al., 2007; Gedikli & Jannach, 2010; Malone, Laubacher, & Dellarocas, 2010; Buder & Schwind, 2012).

Recommender systems support users in both stages of the item-choosing, i.e. decision-making, process (Schafer, Konstan, & Riedl, 2001; Xiao & Benbasat, 2007). They first help users in the initial screening stage to form a *consideration set* from all items, in effect helping users deal with information overload, by offering them a list of items they are likely to find salient, i.e. useful or interesting (*find good items*) or by annotating items e.g. with predictions, ratings, or text to allow users distinguish between desired and undesired items (*annotation in context*) (Schafer, Konstan, & Riedl, 2001; Herlocker et al., 2004; Xiao & Benbasat, 2007). In the second stage, recommender systems support users in comparing and evaluating the items in the consideration set to decide on an item (or items), typically by annotating items with e.g. calculated prediction, ratings or their averages, and textual comments, including tags, comments, and reviews (Resnick & Varian, 1997; Schafer, Konstan, & Riedl, 2001; Herlocker et al., 2004; Xiao & Benbasat, 2007).

Recommender systems can be personalized (e.g. a personalized list of recommendations formed by using a collaborative filtering approach) or non-personalized (e.g. TopN lists) in both stages (Schafer, Konstan, & Riedl, 2001; Lam, Frankowski, & Riedl, 2006; Cremonesi, Garzotto, & Turrin, 2012a). Furthermore, they can be algorithmic (e.g. collaborative filtering computing) or non-algorithmic (e.g. unfiltered customer reviews) (Resnick & Varian, 1997; Lam, Frankowski, & Riedl, 2006).

eWOM is part of recommender systems and covers ratings (both individual ratings and their averages) and textual annotations, including comments, reviews, and tags (when used to rate items or express opinions on items, i.e. mini-review them) (Schafer, Konstan, & Riedl, 2001; Lee, Son, & Han, 2007; Pathak et al., 2010). In addition, eWOM also includes algorithmic and personalized approaches, as user-based collaborative filtering can be seen as mimicking word-of-mouth and providing community opinion (Schafer et al., 2007; Xiao & Benbasat, 2007; Pathak et al., 2010). However, not all algorithmic approaches, e.g. content filtering, can be seen as eWOM, as they are not really about using "the opinions of like-minded people to generate recommendations" (Xiao & Benbasat, 2007). Table 1 contrasts the two types of recommender systems in a somewhat simplified manner.

| Recommender systems | |
|---|---|
| ***Algorithmic recommenders*** | ***Community opinion/eWOM recommenders*** |
| *Examples* | |
| • Lists of recommended items<br>• Prediction calculations | • Lists of top sellers<br>• Lists by editors<br>• Ratings<br>• Reviews<br>• Comments<br>• Tags (partially) |
| *Use of ratings and other machine-readable item evaluations (including implicit evaluations, e.g. viewings)* | |
| Raw material for complex calculations | As-is or raw material for simple calculations (e.g. average of ratings) |
| *Differences* | |
| • Help find (potentially) salient items<br>• Personalized<br>• Algorithmic<br>• Automatic<br>• Not traceable back to individual users in the community<br>• Only positive recommendations | • Help make decisions concerning items<br>• Non-personalized<br>• Non-algorithmic<br>• Manual<br>• Some traceable back to individual users in the community<br>• Also negative recommendations |

**Table 1.** A simplifying summary of recommender systems that contrasts algorithmic recommenders with community opinion/eWOM recommenders (the differences shown here represent *typical* differences between the two types of recommenders rather than absolute differences, and some differences shown here as binary are actually continuums).

# 3 Early Days: First Recommender Systems

Recommender system history started in the early-to-mid 1990s and they developed in parallel with the rapid spread of Internet use (Konstan & Riedl, 2012; Bobadilla, 2013). In this section, we look at some of the first recommender systems that were developed in the research community to see how it all began. However, we focus on the most significant and famous early examples rather than try to provide a complete list of recommender system projects in the academia.

## 3.1 TAPESTRY AND THE BEGINNING OF THE RECOMMENDER SYSTEMS

The first recommender system, Tapestry, was introduced in 1992 by Goldberg et al. (1992) (Resnick & Varian, 1997; Terveen & Hill, 2001; Xiao & Benbasat, 2007). In the same paper, Goldberg et al. (1992) also introduced the concept and term *collaborative filtering* that subsequently became the common name for such approaches that assisted users in the face of a growing information overload (Resnick & Varian, 1997; Maltz & Ehrlich, 1995; Konstan & Riedl, 2012) by utilizing *social strategy* approach, i.e. recommendations based on what “other like-minded, similarly situated people” had done in the past (Hill et al., 1995). However, other names were also suggested in the beginning of the recommender system research, such as *social filtering* (Hill et al., 1995) and *social information filtering* (Shardanand & Maes, 1995). As mentioned, the more general term *recommender system* was coined by Resnick and Varian in 1997 and it subsequently became the de facto term for these systems.

Tapestry, an experimental mail filtering system, was developed at the Xerox Palo Alto Research Center by Goldberg et al. (1992). Its goal was to help users deal with information overload that resulted from the increasing number of email documents being sent. Goldberg et al. (1992) claimed that mailing lists were not overly helpful, as existing lists rarely mapped well to the interests of individual users. They considered user-constructed filters that scanned all the lists to select documents, i.e. content-based filtering, a better approach but felt that the filtering process could be enhanced by involving humans. In effect, knowledge is accumulated as users access messages and appraise them, forming opinions, and the core idea of collaborative filtering was to gather this knowledge and provide suitable access to it for other users (Maltz & Ehrlich, 1995). Consequently, Tapestry was built to support collaborative filtering in addition to content-based filtering (Goldberg et al., 1992).

However, the way Goldberg et al. (1992) used the term *collaborative filtering* differs somewhat from its later use; they defined collaborative filtering as meaning that "people collaborate to help each other perform filtering by recording their reactions to documents they read." Goldberg et al. (1992) called these *free-text reactions* (Hill et al., 1995), e.g. that the document was particularly interesting or uninteresting, *annotations*. The annotations could be accessed in Tapestry with filters: A user sets a filter that checks annotations, and a document matching the filter is returned when the specified annotation is added to the document by a user (Goldberg et al., 1992). These filters were running continuously[6] (Goldberg et al., 1992).

Collaborative filtering approach meant that, unlike with content-based filtering, the filters could not be applied the moment a message arrived because it first had to be read and annotated by somebody before filtering could take place (Goldberg et al., 1992). Consequently, Goldberg et al. (1992) envisioned two types of users, eager ones who read all the documents in a class and more casual users who waited until the eager ones had annotated documents and then got the interesting ones through their filters.

Goldberg et al. (1992) also allowed using other users as kind of moderators in Tapestry, as users could specify wanting those messages that user X had annotated. Moreover, Tapestry allowed using implicit feedback in the filtering (Goldberg et al., 1992). For example, a user could specify a filter that provided them with all the messages in a newsgroup

[6] In effect, Tapestry did not have an algorithm to form neighborhoods of similar users to base automatic recommendations for a user on the preference data from the nearest neighbors of the user, as user-based collaborative filtering would be understood to mean today. Instead, *collaboration* meant that users annotated documents as a community and *filtering* that the annotations were used by the filters that each user set manually for themselves in scanning the documents. In practice, filtering was content-based and the annotations provided the content to scan.

that were replied to by user X *OR* user Y whom they knew to read most if not all messages in the newsgroup. In effect, Tapestry allowed making semantic metadata queries, e.g. written by X *OR* replied by Y (Schafer et al., 2007).

Goldberg et al. (1992) maintain that Tapestry was "more than a mail system, because it is designed to handle any incoming stream of electronic documents", and that was not simply a mail-filtering mechanism as it also functioned as "a repository of mail sent in the past." However, the twin task of writing annotations and specifying filters proved to require significant user effort (Hill et al., 1995), and the vast majority of documents went unannotated (Maltz & Ehrlich, 1995). Still, Tapestry had taken the all-important first step by "incorporating user actions and opinions into a message database and search system" (Schafer et al., 2007).

The Tapestry model of interaction, i.e. leaving the pulling of recommendations out of the database up to the user who wants recommendations, is known as *pull-active collaborative filtering* (Schafer et al., 2007). Very soon after Tapestry, the potential inherent to the human "information hubs" that occur in organizations naturally began to be recognized and exploited more widely (Schafer et al., 2007). For example, Maltz and Ehrlich (1995) used the conceptual notion of collaborative filtering to build inside the Lotus Notes environment a kind of collaborative filtering system that allowed people to send "pointers" to their colleagues on documents that they had found and felt interesting to the receiver. A *pointer* contained "a hypertext link to the document of interest", "contextual information" ("title and date of document, name of source database, name of sender"), "and optional comments by the sender". The motivation of Maltz and Ehrlich (1995) was to support collaboration and information sharing among colleagues—the system was also implemented within the Xerox PARC—and to allow "information mediators", i.e. people good at locating relevant information and using it to solve current problems, easily distribute the information sources they had found. This kind of approach is known as *push-active collaborative filtering*, as a person can push documents to others (Schafer et al., 2007).

## 3.2 AUTOMATED COLLABORATIVE FILTERING

Both Tapestry and especially the system by Maltz and Ehrlich are limited by the fact that they are designed for small workgroups where the members more or less know each other (Maltz & Ehrlich, 1995; Schafer et al., 2007). In effect, pull-active systems necessitate users to know whose opinions to trust while push-active systems necessitate users to know who might be interested in what kind of content (Schafer et al., 2007). *Automated collaborative filtering* systems soon came into the picture to deal with this limitation "by using a database of historical user opinions to automatically match each individual to others with similar opinions" (Schafer et al., 2007).

In fact, the earliest recommender systems, or collaborative filtering systems in the contemporary terminology, were generally designed to give users *information about* items. With automated collaborative filtering systems, collaborative filtering started to be used more behind the scenes to personalize the content to the user. (Schafer et al., 2007)

Early automated collaborative filtering systems included GroupLens (Resnick et al., 1994), Ringo (Shardanand & Maes, 1995), and Bell-core's Video Recommender (Hill et al., 1995). In fact, despite varying terminology due to a common language not having developed yet, all these three systems elicit ratings from users, compute the correlations between pairs of users to identify each user's neighborhood (i.e. most similar users in taste), and combine the ratings of the neighbors to generate recommendations. In other words, all use a basic collaborative filtering approach to generating recommendations. (Schafer et al., 2007)

#### GroupLens: Customizing Predictions in Usenet

GroupLens, the development of which started in 1992 (Konstan et al., 1997), was a collaborative filtering system developed by Resnick et al. (1994) for the Usenet netnews that helped users find articles of interest from the huge stream of articles; the amount of information available meant that nobody could follow it all (Resnick et al., 1994). Moreover, most users felt that the signal-to-noise ratio was too low in addition to which tastes do differ; no one article can appeal to all readers of a newsgroup (Resnick et al., 1994; Konstan et al., 1997). In fact, users in general found only 5–30% of the articles in a newsgroup interesting (Konstan et al., 1997).

If Tapestry was based on the idea that users are better evaluators of articles than computed functions (Goldberg et al., 1992; Maltz & Ehrlich, 1995) as humans, for example, do not share the difficulties of computers when faced with synonymy, polysemy, and context when judging the relevance of a text (Resnick et al., 1994), GroupLens in turn was based on the heuristic that people who agreed in their subjective evaluations in the past are likely to agree again in the future, too (Resnick et al., 1994). The system allowed users to rate articles with numbers from 1 (bad) to 5 (good), and it used these ratings first to correlate the ratings to determine which users' ratings were most similar to each other and then to predict how well users will like new articles based on the ratings that similar users have given them (Resnick et al., 1994). In this way, GroupLens customized the prediction score to each user, in effect accommodating different tastes (Resnick et al., 1994). Furthermore, users did not need to know each other and privacy could be maintained since ratings entered under pseudonyms worked as well as those that were signed (Resnick et al., 1994). However, the need for ratings rendered the system susceptible to cold-start problem (Maltz & Ehrlich, 1995; Konstan et al., 1997).

Resnick et al. (1994) also discussed at length what kinds of social changes the use of collaborative filtering might introduce in a community. They felt that such systems as GroupLens might reduce the need for moderators, as now groups could function as moderators. Furthermore, they felt that GroupLens might effectively do away with the need to split newsgroups into subgroups, as collaborative filtering itself is likely to lead users to articles of interest more effectively. However, the question they asked was whether such peer groups would be permeable or would the global village end up fracturing into tribes. In effect, collaborative filtering might do away with kill-files, as it would filter away uninteresting articles very quickly. Finally, they felt that new social patterns would have to develop to encourage social behavior beneficial to the community, e.g. reviewing articles that had already received some low ratings.

#### Bell-core's Video Recommender

Bell-core's Video Recommender used a similar approach to that of GroupLens to support recommending movies through email and the web in a virtual community of movie buffs (Hill et al., 1995; Konstan & Riedl, 2012). The approach adopted by Hill et al. (1995) was based on the idea that the community could be virtual: Personal relationships are not necessary in social filtering and people can influence each other as though they interacted without actually interacting directly. They showed that their *social filtering* approach—quintessentially collaborative filtering—produced better predictions by far than two nationally known movie critics, i.e. actual user ratings were significantly closer to the predictions the system calculated.

#### Ringo: A Personal Music Recommender

Ringo, a personalized music recommender by Shardanand and Maes (1995) that became available to the public in 1994, was also developed to deal with information overload contemporaneously with GroupLens and Video Recommender. The point of Shardanand and Maes' (1995) *social information filtering* approach—again, quintessentially collaborative filtering (Konstan & Riedl, 2012)—was to automate the process of word-of-mouth recommendations to overcome the limitations of content-based filtering approaches commonly used at the time. Shardanand and Maes (1995) listed the following limitations:

1) Content-based filtering required items to be in machine-parsable form, e.g. text, or the attributes needed to be assigned by hand. However, with media files, e.g. sound, photographs, and video, analyzing items automatically for relevant attribute data was not possible.

2) Serendipitous recommendations could not be generated with content-based filtering.
3) Content-based filtering could not distinguish items based on quality, style, or point-of-view; e.g. a well written and badly written article that used the same terms were seen as equal.

The basic idea of social information filtering was to maintain *user profiles*, i.e. a record of user ratings of music. These profiles were then compared for similarity, and the most similar profiles were used to generate recommendations for or against items to the user. The algorithm with which the similarity was determined could vary. (Shardanand & Maes, 1995)

Again, the idea was to recommend items to a user based on the ratings by other users with similar taste. In this way, the limitations of content-based filtering could be overcome: Items did not need to be machine-parsable, recommendations could be serendipitous as they did not need to be content-wise similar to the items the user had rated, and recommendations were based on quality instead of more objective item properties. (Shardanand & Maes, 1995)

Ringo was accessible through email or the Internet. When a user first accessed Ringo, they were asked to rate 125 artists by how much they liked to listen to them. Users were not required to rate the artists they were not familiar with or the artists on whom they did not have a strong opinion. The rating scale was from 1 ("Pass the earplugs") to 7 ("BOOM! One of my FAVORITE few!"). Ringo also allowed writing short reviews but they were not used in prediction calculating. (Shardanand & Maes, 1995)

Once the profile was initiated, the user could get predictions: Suggestions on new artists/albums that they were likely to like, lists of artists/albums that they were likely to hate [7], and predictions about a specific artist/album. There were also other features, e.g. adding new artists and albums to the database, viewing "Top 30" and "Bottom 30" lists, and possibility of subscribing to a newsletter. In other words, Ringo provided users with both personalized and non-personalized recommendations. (Shardanand & Maes, 1995)

Shardanand and Maes (1995) tested four possible algorithms in Ringo. They set aside 20% of the ratings in each profile to form a *target set* (ratings that they attempted to predict) while the remaining 80% constituted the *source set*. The criteria for evaluation was the *mean absolute error* (needs to be minimized), the *standard deviation of the errors* (also to be minimized),

[7] In effect, Ringo is one of the very few algorithmic recommender systems that also provide *negative* recommendations.

and *T*, or the percentage of target values for which the algorithm can compute predictions[8] (needs to be maximized). The base case against which the algorithms were compared was the average of all ratings received by an artist in the data set (used as a proxy for the predicted score for the artist).

The four algorithms that Shardanand and Maes (1995) evaluated were the mean squared differences algorithm, the Pearson algorithm (Pearson *r* correlation coefficient to measure similarity between profiles), the constrained Pearson *r* algorithm, and the artist-artist algorithm. The first three are what today would be called user-based collaborative filtering and the last represents an item-based collaborative filtering approach[9]. The best algorithm overall, considering both accuracy and the percentage of target values that can be predicted, was the constrained Pearson *r*. Incidentally, Pearson *r* correlation coefficient used by GroupLens was not very efficient according to the tests by Shardanand and Maes (1995).

While the tests showed that Ringo is able to make personalized recommendations, Shardanand and Maes (1995) argued that ultimately such algorithmic tests are not as meaningful as the human response. In the beginning, with fewer users, users felt that Ringo performed poorly (Shardanand & Maes, 1995). However, over time as certain critical mass was being reached, it started to perform so well that users felt it was "unnervingly accurate" (Shardanand & Maes, 1995). In effect, Ringo was susceptible to cold-start problems, as many collaborative filtering approaches are.

Altogether, many aspects of Shardanand and Maes's (1995) work were very modern. For example, their emphasis on the human response, or user experience and adoption of the system, is very much in line with how evaluating recommender systems is viewed today. In fact, we will meet again many concepts that they considered in this early work later in this dissertation.

## 3.3 ENTER E-COMMERCE

As research efforts on recommender systems started to take place in academia, e-commerce sites, led by the pioneering efforts of Amazon.com that went online in 1995, also saw their potential and started to implement them to help their users find items of interest out of the huge number of items available (Pathak et al., 2010; Konstan & Riedl, 2012). In fact, the pace in e-commerce was so fast that already in 1997 Resnick and Varian stated that a "flurry of commercial ventures have recently introduced recommender systems for products ranging from Web URLs to music,

[8] Today, this metric would be called *coverage*; see Section 8.1 (discussion on Coverage).

[9] For details on item-based and user-based collaborative filtering, see Section 5.2.

videos, and books" while two years later, in 1999, Schafer, Konstan, and Riedl maintained that recommender systems were "changing from novelties used by a few E-commerce sites, to serious business tools that are re-shaping the world of E-commerce." Today, of course, "recommender systems are embedded in a wide range of commerce and content applications (both online and offline)" (Konstan & Riedl, 2012), and the scope of recommender systems has expanded from collaborative filtering to numerous approaches (Konstan & Riedl, 2012; Bobadilla et al., 2013).

While academic research focused on evaluating algorithms in statistical ways, testing the accuracy of predictions in various ways, businesses focused from the start on the user, or customer, experience and adapted their recommender approaches to fit with their sales and marketing efforts (Xiao & Benbasat, 2007; Konstan & Riedl, 2012). Only recently has the research community begun to widely appreciate the importance of the user experience to the success or failure of recommender systems (Konstan & Riedl, 2012) even though some early work, e.g. Shardanand and Maes (1995), had already emphasized it.

Moreover, online businesses also moved away from the user-based collaborative filtering, used e.g. in GroupLens and Video Recommender, since they had to deal with hundreds of thousands or even millions of users, and user-based (or user-user) approaches are too slow for real-time use under such conditions (Linden, Smith, & York, 2003; Konstan & Riedl, 2012). Instead, online retailers such as Amazon.com opted for the item-based (or item-item) approaches that have faster response times, especially if the item relationships are pre-computed offline as relationships between items tend to be rather stable, and that in fact tend to be slightly more accurate, too (Linden, Smith, & York, 2003; Konstan & Riedl, 2012). Interestingly, researchers working on Ringo did also test one item-based algorithm in addition to three user-based ones, but performance was not a consideration for them as the system only had about two thousand users at the time (Shardanand & Maes, 1995). In effect, in e-commerce the underlying algorithms quickly "converged to a few particularly useful ones" (Konstan & Riedl, 2012).

Today, the scope of e-commerce also includes such social media services as Twitter[10], LinkedIn[11] and Facebook[12]. Recommender systems play an important role there, too, as such services use them in multiple ways, e.g. recommending news in addition to recommending potential friends (Konstan, 2012). The importance of recommenders to social media services is underlined by the fact that by 2010, the lack of recommendation service

[10] https://twitter.com/
[11] https://www.linkedin.com/
[12] https://www.facebook.com/

in Twitter "was viewed both internally and externally as an important missing product feature," especially since such sites as LinkedIn and Facebook already offered such functionalities (Gupta et al., 2010). After all, a social media service depends on connecting users, and not assisting users in it with a recommendation service was seen as creating "an excessive burden on users" (Gupta et al., 2013). Naturally, such social media services as Facebook and Twitter also function as platforms for a lot of recommending and sharing activity, too (Gupte et al., 2009; Lerman & Ghosh, 2010). (For further discussion on the role of social media in news recommending activity, see Section 10.3.)

# 4 Benefits of Recommender Systems

Recommender systems can be seen as being beneficial both to service providers and users (Cosley et al., 2003; Campochiaro et al., 2009; Pathak et al., 2010; Pu, Chen, & Hu, 2011; Cremonesi, Garzotto, & Turrin, 2012a). For users, recommender systems lower the transaction costs of finding and selecting items, be it in an online shopping environment or library (Neumann, 2007; Hu, Liu, & Zhang, 2008). Moreover, recommender systems have been shown to improve decision quality (Xiao & Benbasat, 2007; Pathak et al., 2010). For e-commerce players, recommender systems can enhance revenues, as they are increasingly seen as effective means to sell more products (Schafer, Konstan, & Riedl, 2001; Campochiaro et al., 2009; Castagnos, Jones, & Pu, 2009; Pu, Chen, & Hu, 2011). For other online establishments that are not driven by commerce, recommender systems are no less attractive (Neumann, 2007). For scientific libraries, for example, recommender systems offer user support possibilities that can enable them to become digital information centers by allowing them to move beyond catalog searches and to support different user needs (Neumann, 2007).

## 4.1 Benefits to Users

In discussing how recommender systems are defined, we already discussed the two-stage decision making process wherein recommender systems support the user (see Section 2.4). In effect, recommender systems suggest items to users and then help users decide which item(s) to select for purchasing, viewing, downloading, and so on (Schafer, Konstan, & Riedl, 2001; Svensson, Höök, & Cöster, 2005).

An important function of recommender systems in this process is to provide users with *confidence* to act, be it to purchase or some other action to be taken on an item (Castagnos, Jones, & Pu, 2009; Pu, Chen, & Hu, 2011). In this sense, recommender systems are also *persuasive* systems and have a persuasive function in the decision making process (Gretzel & Fesenmaier, 2006; Knijnenburg, Reijmer, & Willemsen, 2011; Buder & Schwind, 2012; Cremonesi, Garzotto, & Turrin, 2012a).

### Lowering Transaction Costs

If we look at how recommender systems assist their users from the viewpoint of the theory of transaction cost economics, the purpose of recommender systems is to reduce transaction costs, i.e. to enable users to choose the item with the lowest overall transaction cost (Neumann, 2007; Hu, Liu, & Zhang, 2008). Transaction costs are thought to occur because decision making is subject to limited cognitive processing power and not all possible scenarios can be considered, in part due to bounded rationality[13] (Hu, Liu, & Zhang, 2008).

In order to find and select an item, a user must go through a *transaction process* that starts with a search for relevant items and is followed by a comparison stage that involves in e-commerce e.g. price comparisons and item quality evaluations, and questions related to ordering, delivering, and post-sales services, depending on the nature of the item. The process can involve product, process, and psychological uncertainties (especially in the case of experience products, as the quality of such products can only be understood after trying them). *Product uncertainty* refers to the possibility of a consumer finding out after consuming the product (e.g. reading a book) that what they received is different from what they perceived when ordering it. *Process uncertainty* in turn refers to consumers purchasing products from undesired vendors. Finally, *psychological uncertainty* refers to the emotional cost of uncertainty. Taken together, *uncertainty* refers to the costs connected to unexpected outcomes due to information asymmetry. (Hu, Liu, & Zhang, 2008)

High uncertainty means high transaction cost, and the natural inclination of the consumer is to judge the intrinsic quality of the product based on all the available information and to choose the item with the lowest transaction cost, i.e. with lowest uncertainty. *Uncertainty reduction theory* states that when faced with uncertainty, the consumer will attempt to reduce uncertainty to eliminate or at least mitigate risks brought on by uncertainty to maximize the outcome value. Uncertainty reduction theory

[13] *Bounded rationality*, a field pioneered by Herbert Simon and widely studied by Daniel Kahneman, Amos Tversky, and their collaborators (Rubinstein, 1998, pp. 16, 187), refers to the fact that rationality of decision making is bounded, i.e. limited, by various factors, such as limited information, cognitive limitations inherent to the human mind, and limited time available for decision making (Simon, 1955, 100-101, 110-113). See also: http://en.wikipedia.org/wiki/Bounded_rationality.

offers a framework for understanding the use of recommender systems, especially eWOM type of systems such as customer reviews, to infer product quality, to reduce (product) uncertainty, and to make the final choice of item. (Hu, Liu, & Zhang, 2008)

Recommendation systems are able to lower transaction costs (Hu, Liu, & Zhang, 2008; Pathak et al., 2010). They can help us efficiently locate items of interest, thus reducing the number of items to browse, e.g. with personalized lists of salient items, and they can reduce uncertainty about the products, e.g. with customer reviews, thereby also reducing the overall time it takes to make a decision (Schafer, Konstan, & Riedl, 2001; Xiao & Benbasat, 2007; Hu, Liu, & Zhang, 2008; Campochiaro et al., 2009; Pathak et al., 2010). In other words, recommender systems assist us in cognitively demanding tasks, e.g. by reducing the complexity of searching for items of interest with item-suggestions (Schafer, Konstan, & Riedl, 2001; Xiao & Benbasat, 2007; Pommeranz et al., 2012). In addition, they assist in decision making by providing product information (both personalized and non-personalized), summarizing community opinion, and providing community critiques (Schafer, Konstan, & Riedl, 2001). As a result, recommender systems have been shown to improve decision quality (Xiao & Benbasat, 2007; Pathak et al., 2010).

Moreover, in addition to suggestions and predictions, recommender systems offer users information and even education (Konstan, 2008). In fact, besides being useful, using recommender systems can and should be fun and enjoyable (McNee, Riedl, & Konstan, 2006a).

Recommender systems are user-centric by nature as customer value is the business driver and they provide value-added services to users (Pathak et al., 2010). In fact, recommender systems “are becoming essential tools for people to deal with information overload, huge search spaces and complex choice sets in different domains” (Pommeranz et al., 2012). Their importance to users can also be seen in that in 2008, a Forrester survey showed that 64 percent of the respondents wanted to see user ratings and reviews on web sites, significantly more than e.g. those who wanted personalization (37%) (Kee, 2008). Also, according to the report, 68 percent of the respondents read at least four reviews before buying something while almost a quarter read at least eight before making a purchasing decision (Kee, 2008).

### Personalization

Recommender systems are at the epicenter of personalization, the drive to have websites adapt to each user and provide personalized user experience (Schafer, Konstan, & Riedl, 2001; Linden, Smith, & York, 2003; Kumar & Benbasat, 2006; Anand & Mobasher, 2007). For example, it is virtually impossible for two users of Amazon.com to see exactly the same content over any longer visit to the site, as personalization begins from the first choice the user makes; e.g. the books that are recommended to us are based on items that we and other customers have viewed and bought (Linden, Smith, & York, 2003; Kumar & Benbasat, 2006; Anand & Mobasher, 2007; Brusilovsky, 2007).

As discussed, according to the narrow definition of recommender systems, personalization is an integral part of any recommender system, e.g. Pommeranz et al. (2012) define recommender systems as "tools that provide personalized recommendations to people" and Knijnenburg et al. (2012) state that recommender systems "offer each user a personalized subset of items, tailored to the user's preferences."

In contrast, Schafer, Konstan, and Riedl (2001) speak of varying degrees of personalization rather than dichotomy between personalized and non-personalized recommendations. Konstan (2008) defines four levels of personalization:

1. *Generic*: Everybody receives the same recommendations. This is quintessentially non-personal level of recommending. (Schafer, Konstan, & Riedl, 2001; Konstan, 2008)

2. *Demographic*: All members of the target group receive the same recommendations (Konstan, 2008).

3. *Ephemeral*: Recommendations match the current activity of the user; recommendations respond to user's navigation and item-selection. An example of this is Amazon.com's *Customers who…* recommendations. (Schafer, Konstan, & Riedl, 2001; Konstan, 2008)

4. *Persistent*: Recommendations match long-term interests of the user (Konstan, 2008). Persistence requires users to maintain persistent identities in the system but also rewards them with most deeply personalized recommendations (Schafer, Konstan, & Riedl, 2001).

While a clear and useful tool for conceptualization, Konstan's (2008) four levels naturally represent a kind of a functional simplification, as personalization is, in fact, a continuum across several dimensions (Schafer, Konstan, & Riedl, 2001).

However, while the levels are numbered from one to four, the number signifies only the level of personalization, not any sort of order of preference—more personal does not mean better. In fact, simple, non-personalized recommendations are well received by users even though users are aware of their low utility (Cremonesi, Garzotto, & Turrin, 2012a). The user type may affect the equation, as Knijnenburg, Reijmer, and Willemsen (2011) found that especially novices and maximizers, i.e. users who aspire to the best possible result, benefit more from non-personalized recommenders that simply display the most popular items (TopN).

Svensson, Höök, and Cöster (2005) suggest that in the domains where most items are of interest to users and most items are of approximately the same quality, a recommender that simply shows the most popular items may work well; little personalization is necessary. Similarly, Recker, Walker, and Lawless (2003) point out that if all users are in general agreement about the quality of items, personalized recommendations can add little value. In the same vein, Schafer, Konstan, and Riedl (2001) state that recommenders can efficiently summarize community opinion, e.g. in form of aggregate or summary ratings, when personalization is either impractical or unnecessary.

However, Svensson, Höök, and Cöster (2005) studied a social navigation system focusing on food recipes (over 3,000 recipes) with 598 users while Recker, Walker, and Lawless (2003) focused on an e-learning setting where the numbers of users and items are also low. The conditions they discuss are much more likely to occur in such settings than on Amazon.com with millions of users and items. This simply underlines the huge variety of settings where recommender systems are used; it would be unrealistic to expect exactly the same contextual factors in all or that exactly the same approach would work in all settings.

Also, not all types of recommenders lend themselves easily to personalization. For example, customer reviews have proven resilient to attempts to personalize them (Ludford, 2007). Such recommendations remain manual in the sense that users need to read them and draw their own conclusions about item (Lam, Frankowski, & Riedl, 2006).

Personalization is very much data-driven (Anand & Mobasher, 2007; Cremonesi, Garzotto, & Turrin, 2012b; Jeckmans et al., 2013). Recommender systems depend on user models to provide personalized recommendations (Konstan & Riedl, 2012). Before the model is formed, personalized recommendations cannot be given (Cremonesi, Garzotto, & Turrin, 2012b), and, other things being equal, the more detailed information there is of the user, the more accurate the recommendations can be (Schafer et al., 2007; Jeckmans et al., 2013). Data sparsity—a state where there is not enough overlap in ratings across users, making it difficult to find other users with whom to correlate the current user (Konstan et al., 1997; Burke, 2002)—is another problem that can make personalizing recommendations challenging (Schafer et al., 2007). Before enough data exists to fire up the user model, many commercial systems simply fall back to lower levels of personalization, e.g. contextual, demographic, and most popular (TopN) approaches (Konstan & Riedl, 2012). The resulting relentless drive to collect as much data as possible leads to mounting concerns about privacy (Resnick & Varian, 1997; Kobsa, 2007; Anand & Mobasher, 2007; Schafer et al., 2007: Jeckmans et al., 2013).

Interestingly, the degree of perceived personalization directly affects cognitive and emotional trust; emotional trust mediates the impact of cognitive trust on the intent to adopt the recommender system as a delegated agent fully and on the intent to adopt it as a decision aid partially (Komiak & Benbasat, 2006). Therefore, personalization can also affect whether or not the user will use the system and trust its suggestions.

Researchers envision many areas where they feel that personalization in recommender systems should move. Knijnenburg, Reijmer, and Willemsen (2011) see that recommender systems should adapt their interaction method to the user, as different types of users appear to benefit from different types of recommendations. Harper et al. (2005) feels that we should consider personalizing interfaces in addition to personalized content. In the same vein, Cleger-Tamayo, Fernández-Luna, and Huete (2012) see recommender systems as providing possibilities to offer personalized media front pages in the area of news recommending. Schafer et al. (2007) feel that personalization should also apply to a group of users; recommendations to a group should be personalized to the group. Perhaps the most revolutionary is Hangartner's (2007) idea that we can move beyond focusing on "delivering an isolating, customized experience to a person" to "connecting an individual with affinity communities who can provide information of value to that individual."

## 4.2 Benefits to Service Providers

In the discussion about the benefits of recommender systems to the service provider, we focus on e-commerce, as that is the area that has received the most attention. We touch upon the benefits to news providers and e-learning environments when discussing those areas, and omit discussing them here. Suffice to say here, while the point for many non-commercial areas are the benefits to the user, i.e. reduced transaction costs (Hu, Liu, & Zhang, 2008), these benefits do often tie to the benefits and aspirations of the non-commercial providers, e.g. the aspirations of scientific libraries wanting to develop into digital information centers (Neumann, 2007).

Obviously, the main interest of e-commerce establishments is to sell products, and recommender services are increasingly recognized as effective means to this end, as employing a recommender system can affect both the *lift factor* (sales volume increase due to introducing a recommender system) and the *conversion rate* (the ratio between the number of sales made due to recommendations and the number of recommendations) (Castagnos, Jones, & Pu, 2009; Cremonesi, Garzotto, & Turrin, 2012a).

With too numerous products to browse, interactive decision aids such as recommender systems are becoming indispensable for online shops in assisting customers in finding and selecting products (Castagnos, Jones, & Pu, 2009). Not only do recommender systems assist customers in finding salient products, providing efficient product search tools and ideally serendipitous recommendations, but they also provide them with the necessary confidence to purchase product(s) (Castagnos, Jones, & Pu, 2009; Pathak et al., 2010; Pu, Chen, & Hu, 2011; Konstan & Riedl, 2012). Recommender systems tend to diversify purchases, and, in fact, their effect appears more influential when the items found are less obvious, and so recommender systems are seen as reinforcing the long-tail trend in e-commerce (Pathak et al., 2010; Cremonesi, Garzotto, & Turrin, 2012a). Consequently, it is no wonder that e-businesses immediately recognized the need to focus on the customer experience rather than algorithmic accuracy (Konstan & Riedl, 2012).

Schafer, Konstan, and Riedl (2001) list three ways that e-commerce sales are enhanced by recommender systems:

1) *Converting browsers into buyers*. Recommender systems can help convert browsers into buyers by helping them find products that interest them from the multitude of products and they can instill the necessary confidence necessary for purchasing, as they can provide signals of quality ad fit (Schafer, Konstan, & Riedl, 2001; Castagnos, Jones, & Pu, 2009; Pathak et al., 2010). In fact, recommendations tend to be viewed as the correct answer, as consumers appear to have some inherent trust on recommender systems (Adomavicius et al., 2013).
2) *Increasing cross-sell*. Recommender systems are especially apt in suggesting additional or related products for customers, and if recommendations are good, the result should be increased sales (Schafer, Konstan, & Riedl, 2001). While the items that users have placed in the shopping cart provide a basis for cross-selling opportunities as they indicate the current interest of the customer (Schafer, Konstan, & Riedl, 2001), signaling and advertisement effects also help in cross-selling (Pathak et al., 2010). Signaling refers to the prior purchases of other customers having signaling effects while recommendations also function as advertisements, tending especially to increase awareness of relatively obscure items, i.e. enhancing the long-tail sales (Pathak et al., 2010; Cremonesi, Garzotto, & Turrin, 2012a).
3) *Building loyalty*. When the competitor is just a click away, gaining and maintaining customer loyalty is essential (Schafer, Konstan, & Riedl, 2001). Loyalty can be improved by creating a value-added relationship between the customer and the e-commerce site (Schafer, Konstan, & Riedl, 2001). Collecting customer data and

operationalizing it as recommendations that match the customer interests leads to loyalty, to customers returning to the site (Schafer, Konstan, & Riedl, 2001; Pathak et al., 2010; Pu, Chen, & Hu, 2011). Moreover, recommender systems, once they have sufficient data to make matching recommendations, increase switching costs, as customers would have to invest time on a competing site before it could match the fittingness of the recommendations even if it offered exactly the same recommendation systems (Schafer, Konstan, & Riedl, 2001; Pathak et al., 2010). Customer satisfaction also increases the likelihood of customers recommending the site to their friends and family (Pu, Chen, & Hu, 2011). Finally, recommender systems can also help build a community of customers around products (Konstan, 2008), further fostering loyalty.

In fact, Castagnos, Jones, and Pu (2009) go so far as to suggest that "sites that do not employ intelligent tools will not only see poor purchase volumes but also experience less traffic because consumers are more likely to return to a site employing recommender systems", claiming that adopting the correct tools can affect a store's very survival. In effect, a good recommender system that matches customers with salient products can inspire trust in customers (Cosley et al., 2003). In contrast, Ozok, Fan, and Norcio (2010) claim that having bad recommender systems does not result in e-commerce sites to be seen as bad places for online shopping, i.e. recommender systems do not influence the overall opinion of the consumer concerning the shopping site.

In addition to the three ways that recommender systems enhance e-commerce sales, Schafer, Konstan, and Riedl (2001) also identify "the five business goals and the application models used to address them."

1) *Helping new and infrequent visitors: Broad recommendation lists*. Broad, non-personalized recommendations, such as "best sellers, best sellers in a category, editor and expert recommendations, and other collections of products selected either manually or through simple statistical summarization", work well to engage a new or infrequent visitor (Schafer, Konstan, & Riedl, 2001). As little is known about these visitors, broad lists that require low level of input are effective (Schafer, Konstan, & Riedl, 2001). On the other hand, little personalization is possible and personalization level is limited to ephemeral contextual information, e.g. what product the visitor is currently looking at (Schafer, Konstan, & Riedl, 2001). Still, engaging customers this way is not a major problem, as non-personalized lists are well received by them (Cremonesi, Garzotto, & Turrin, 2012a). Also, broad recommendations allow e-commerce sites "to adjust pricing and inventory to match the

recommendations since they can be assured that these recommendations will reach a large audience" (Schafer, Konstan, & Riedl, 2001; see also Konstan & Riedl, 2012).

2) *Building credibility through community: Customer comments and ratings*. As customers often feel that e-commerce sites are only geared at making sales, such sites need ways to enhance credibility. Customer reviews, comments, and ratings offer a way to provide a high degree of credibility in addition to creating a sense of community. (Schafer, Konstan, & Riedl, 2001)
3) *Inviting customers back: Notification services*. E-commerce sites that know the interests of their customers can leverage this information to invite customers back by informing them of new products of interest or when products of interest are on discount. (Schafer, Konstan, & Riedl, 2001)
4) *Cross-selling: Product-associated recommendations*. As discussed above, knowing the interests of the customer, either from current ephemeral context or from user model data, provides an effective way to suggest items. These recommendations are well suited to be integrated into a product information page, e.g. Amazon.com's *Customers who…* recommendations. (Schafer, Konstan, & Riedl, 2001)
5) *Building long-term relationships: Deep personalization*. Deep personalization is based on wide amount of user model data that allows generating persistent personalized suggestions or predictions (Schafer, Konstan, & Riedl, 2001). Long-term relationship is the goal of most e-commerce, and deep personalization is seen as a way to provide value to the customer and to erect greater competitive barriers by increasing switching cost (Schafer, Konstan, & Riedl, 2001; Pathak et al., 2010). There are numerous algorithmic approaches to deep personalization (Schafer, Konstan, & Riedl, 2001).

### Persuasion

Importantly for increasing sales, recommender systems are, in essence, persuasive systems in that they try to persuade users to follow recommendations, i.e. induce behavioral changes (Gretzel & Fesenmaier, 2006; Knijnenburg, Reijmer, & Willemsen, 2011; Buder & Schwind, 2012; Cremonesi, Garzotto, & Turrin, 2012a). In addition to behavior, recommender systems can influence the user's attitudes, beliefs, and decisions (Cremonesi, Garzotto, & Turrin, 2012a). As mentioned above, recommender systems do induce behavioral changes by enhancing the lift factor and the conversion rate, and by diversifying purchases and orienting them towards the long tail, i.e. the less obvious choices (Cremonesi, Garzotto, & Turrin, 2012a). In fact, introducing recommender systems increases overall sales (Castagnos, Jones, & Pu, 2009; Cremonesi, Garzotto, & Turrin, 2012a).

It appears, rather logically, that persuasiveness of a recommender system depends greatly on the trust that it inspires in the user, as users continue to use only recommender systems that they trust (Komiak & Benbasat, 2006; Knijnenburg, Reijmer, & Willemsen, 2011; Adomavicius et al., 2013). There appear to be social factors in play in this, as the influence that persuasive technologies exert can be social—in effect, users perceive recommender systems as social actors and ascribe them a personality (McNee, Riedl, & Konstan, 2006b; Buder & Schwind, 2012). After all, a recommender system, in a sense, mimics a knowledgeable person who knows the items and the preferences of the user (Buder & Schwind, 2012). Komiak and Benbasat (2006) suggest that users can identify with a recommender system that they perceive as personalized. In effect, in part thanks to the recommender being able to make fitting recommendations, the user can develop a sense of "we-ness" with the recommender system, and identification with a group increases both the "propensity to confer trust" and the "willingness to engage in trusting behavior" (Komiak & Benbasat, 2006). In effect, computers are seen as social actors, and this brings social dimensions into play (Buder & Schwind, 2012).

In addition, recommendations are inherently persuasive in that they are viewed as a "correct" answer by default; users "have some inherent trust in recommender systems" (Adomavicius et al., 2013).

Research on persuasion in recommender systems has mainly concentrated on the persuasiveness of the recommendations. However, it appears that the preference elicitation process actually has a great impact also on how users also perceive the recommendations, as it creates expectations concerning the quality of the recommendations that the system subsequently provides. Gretzel and Fesenmaier (2006) see the elicitation process as a social process, as a form of discourse, that affects how the recommender system and the recommendations it provides are seen. They found that asking more questions only marginally decreases perceived enjoyment but increases perceived value. However, the greater effort also leads to greater expectations, and if the suggestions generated do not match user needs after the user has gone through a great trouble to provide preferences, the suggestions can be seen in a less favorable light than if the user had only gone through a small effort. (Gretzel & Fesenmaier, 2006)

Thus, limiting persuasion research only to the presentation of recommendations and their effects misses the overall picture and can lead to conflicting results that other factors could explain.

# 5 Algorithmic Approaches to Recommender Systems

Various taxonomies have been suggested for recommender systems, as they can be classified along multiple dimensions (Burke, 2002; Thor, Golovin, & Rahm, 2005; Cremonesi, Garzotto, & Turrin, 2012a). Many if not most (e.g. Burke, 2007; Pu, Chen, & Hu, 2012; Bobadilla et al., 2013; Jeckmans et al., 2013) have suggested taxonomies that follow the narrow definition of recommender systems, leaving eWOM and occasionally even non-personalized recommendations outside of the proposed taxonomy. In fact, only the relatively early taxonomy by Schafer, Konstan, and Riedl (2001) appears to cover at least most if not all of the recommendation approaches that our wider definition of recommender systems encompasses.

Focusing on e-commerce recommender applications, the taxonomy by Schafer, Konstan, and Riedl (2001) separates recommender attributes into three categories, 1) the functional I/O, 2) the recommendation method, and 3) other design issues, noting that the categories are not independent. *Functional I/O* is concerned with how the data flows into and out of the systems, as all recommender systems take in some input[14], e.g. preference data, attribute data, and textual inputs, and outputs recommendations that vary e.g. in type, quantity, and look. However, as we will look at the

[14] In fact, one typical way to divide recommender systems into two categories is based on how they gather the preference data to build user preference profiles (Pu, Chen, & Hu, 2012). The recommender systems that are based on explicitly stated preferences can be called *preference-based recommenders* (Pu, Chen, & Hu, 2012) or *explicit recommender systems* (Neumann, 2007), while the recommender systems that use implicitly gathered behavior data, e.g. items viewed, bought, or lent, are called *behavior-based recommenders* (Pu, Chen, & Hu, 2012) or *implicit recommender systems* (Neumann, 2007).

preference elicitation and recommendation output issues elsewhere in this dissertation, we do not get into the details of functional I/O category here. The same goes for *other design issues* category that includes *degree of personalization* and *delivery*.

In this section, we are mainly concerned with the *recommendation method* category. The simplest is the *raw retrieval*, or null recommender, a system that provides a search interface that can be used to query a database; e.g. a user asks a music site for albums by "The Rolling Stones" and the system returns a list of their albums (Schafer, Konstan, & Riedl, 2001). As discussed, however, not all researchers even consider such systems recommender systems; see e.g. Burke (2007). Be that as it may, raw retrieval systems abound in e-commerce sites (Schafer, Konstan, & Riedl, 2001).

Some recommendations are *manually selected* by editors or critics, for example, and provide lists of recommended items to the community. Often, such recommendations are accompanied by a text to help others appraise and make sense of the recommendation. In a sense, these recommendations place *personality* over *personalization*. In these recommendations, no computation is involved, i.e. they are non-algorithmic. (Schafer, Konstan, & Riedl, 2001)

If personalized recommendations are impractical or unnecessary, recommender systems can effectively provide *community opinion* in the form of e.g. *statistical summaries*, including *within-community popularity measures* (e.g. percentage of customers who purchase an item) and *aggregate/summary ratings* (e.g. average rating for an item). Such summaries provide non-personalized recommendations and are easy to compute; they are algorithmic in a sense, but the algorithms are very simple. (Schafer, Konstan, & Riedl, 2001)

After considering the above recommendation methods, Schafer, Konstan, and Riedl (2001) go on to consider algorithmic approaches for generating recommendations, namely *attribute-based*, *item-to-item*, and *user-to-user* approaches. However, the list of algorithmic approaches to automatically generating recommendations has grown significantly into "a plethora of different recommender algorithms" (Said et al., 2012; see also Thor Golovin, & Rahm, 2005; Cremonesi, Garzotto, & Turrin, 2012a). The growth is in part motivated by the fact that no single approach can be the best in all contexts—the "best" approach depends on various contextual factors in addition to algorithmic factors (Thor, Golovin, & Rahm, 2005). Moreover, all algorithmic approaches have their own advantages and disadvantages (Burke, 2002), and understanding where they work the best and where they have problems is essential for deciding which approach to employ (Thor, Golovin, & Rahm, 2005; Schafer et al., 2007). However, surprisingly little information is available on the relative quality of different approaches to guide developers (Thor, Golovin, & Rahm, 2005).

Different types of approaches result in different decision-making processes and outcomes in addition to resulting in different user evaluations of the systems (Xiao & Benbasat, 2007). Moreover, the expertise level of the users affects how they view different approaches. In fact, Xiao and Benbasat (2007) go as far as to suggest that multiple recommender systems be provided, costs permitting, to allow different users to choose the type of recommender to assist them.

There are also many ways to categorize this multitude of approaches. Cremonesi, Garzotto, and Turrin (2012a) see distinguishing between *personalized* and *non-personalized* recommendations as the "key taxonomic criterion". Simply put, non-personalized recommender systems suggest items without considering the user profile, while personalized recommender systems make the recommendations based on the user model.

Recommender system research, especially by those who define recommender systems narrowly, does not typically address non-personalized recommender systems, focusing instead on personalized systems. While personalized systems are defined by various aspects, e.g. their use domain, the data and knowledge sources they use, and how they assemble and present the recommendations, they are typically categorized by the recommendation algorithm they use (Cremonesi, Garzotto, & Turrin, 2012a), i.e. how recommendations are algorithmically identified, or mined and filtered, from the underlying data (Schafer, Konstan, & Riedl, 2001; Cremonesi, Garzotto, & Turrin, 2012a; Park et al., 2012; Bobadilla et al., 2013).

Some recent papers, e.g. Campochiaro et al. (2009) and Park et al. (2012), still simply divide recommender algorithms into two categories, namely *collaborative filtering* and *content-based filtering*. Adomavicius and Tuzhilin (2005) in their influential categorization added to these two hybrid algorithms, i.e. algorithms that combine collaborative and content-based filtering. While these two approaches, collaborative and content-based filtering, are widely used and probably represent the majority of algorithms in use (Burke, 2002; Linden, Smith, & York, 2003; Cremonesi et al., 2011), such binary categorization has long since lost its ability to describe various algorithmic approaches.

The taxonomy by Burke (2007) "that has become a classical way of distinguishing between recommender systems" "and of referring to them" (Cremonesi, Garzotto, & Turrin, 2012a) divided recommendation approaches into five basic techniques: Content-based, collaborative, demographic, utility-based and knowledge-based (Burke, 2002, 2007)[15]. Jeckmans et al. (2013) also follow Burke (2007) and consider collaborative, content-based, demographic, and knowledge-based filtering approaches as the basic recommender types.

Burke (2007) states that all recommender systems employ at least one of these basic techniques that "have complementary advantages and disadvantages." The final category of recommender system algorithms is, in fact, hybrid recommender systems, that combine multiple techniques to achieve synergy and avoid the weaknesses that each individually has (Burke, 2007). Jeckmans et al. (2013) calls recommender types that build on basic types as *improved* recommender types. However, the term *hybrid* appears more popular and is therefore adopted here. Jeckmans et al.'s (2013) terminology is mentioned here simply to demonstrate and underline that even today—their paper is published in 2013—there is no systematic and consistent way of naming the filtering algorithms used in recommender systems.

Given the various categorization schemes and the fact that algorithmic approaches are not in the crux of this dissertation, we briefly discuss below several categories of filtering algorithms that have been discussed in the literature without claiming that it represents any kind of be-all-end-all of such categorizations. Nevertheless, the discussion is by breadth more encompassing than most discussions in the literature. Moreover, we keep the discussion at a general level rather than getting into statistical intricacies of the algorithms.

Before looking at popularity-based, collaborative filtering, content-based filtering, clustering, demographic, case-based (including utility-based, knowledge-based, and critique-based), social, personality-based, context-aware, and hybrid systems, let us glance briefly at the four basic recommendation approaches identified by Burke's (2007) influential work; in the end, while we look here at other approaches in addition to them, they are the most commonly used and discussed approaches to filtering data to generate recommendations.

Burke (2007) states that recommendation techniques can be distinguished based on the knowledge source, i.e. where the knowledge needed to generate the recommendation in the first place comes from. Such knowledge can be the knowledge of user preferences or "ontological or inferential knowledge about the domain, added by a human knowledge engineer" (Burke, 2007). Figure 1 summarizes where the knowledge comes for the four basic types of recommender algorithms.

[15] In fact, utility-based techniques were not discussed in Burke (2007) but were discussed in his earlier work, Burke (2002).

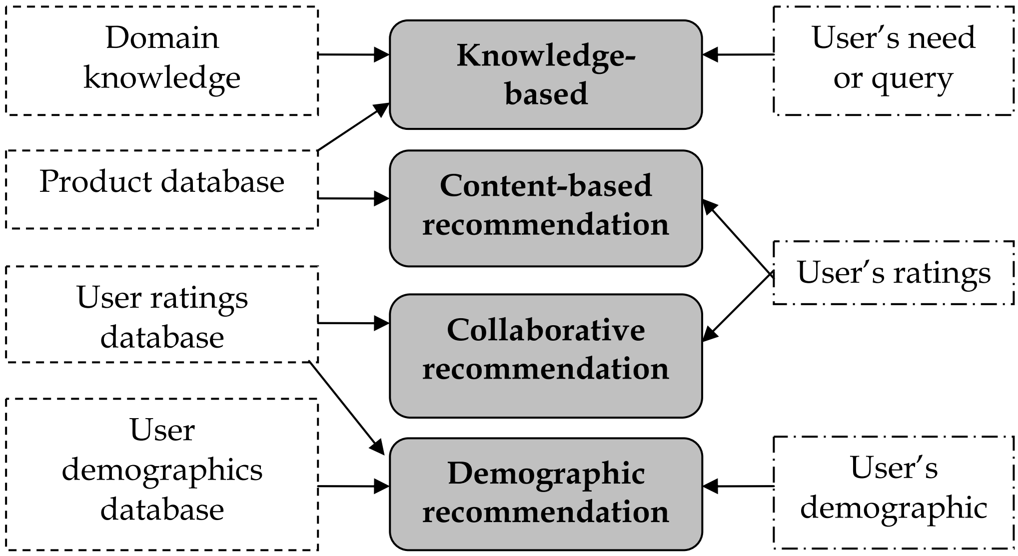


**Figure 1**. Recommendations and their knowledge sources (Burke, 20007).

As all the algorithmic approaches in Figure 1 are discussed below in more detail, the reader is asked to refer back to this figure as necessary in the relevant sections.

## 5.1 POPULARITY-BASED RECOMMENDER SYSTEMS

Popularity-based recommender systems are typically non-personalized and use e.g. *within-community popularity measures* and *aggregate/summary ratings* to recommend the most popular items to users (Schafer, Konstan, & Riedl, 2001; Wei, Shaw, & Easley, 2002). Popularity-based recommender systems underline that recommender systems that provide personalized recommendations form only a subset of recommender systems (Wei, Shaw, & Easley, 2002; Ozok, Fan, & Norcio, 2010) and were consequently added to this list even though the computation needed is very light (Schafer, Konstan, & Riedl, 2001). Popularity-based recommendations, e.g. TopN, are common and well received by users (Schafer, Konstan, & Riedl, 2001; Cremonesi, Garzotto, & Turrin, 2012a).

## 5.2 COLLABORATIVE FILTERING

As discussed, the collaborative filtering approach is seen as mimicking word-of-mouth recommendations both in the sense of them being modeled after traditional word-of-mouth recommendations and in the sense of them using opinions of like-minded people as the basis for the recommendations (Resnick & Varian, 1997; Schafer et al., 2007; Xiao & Benbasat, 2007). However, in a sense, collaborative filtering allows us to advance beyond the limitations of traditional word-of-mouth, as it allows us to use thousands or even millions of opinions (Schafer et al., 2007; Konstan & Riedl, 2012). The approach, in effect, allows us not just to get the opinion of the community-at-large but "a truly personalized view" on an item (Schafer et al., 2007).

Collaborative filtering algorithms are typically divided into *user-to-user* (AKA users-user or user-based) and *item-to-item* (AKA item-item or item-based) collaborative filtering algorithms (Schafer, Konstan, & Riedl, 1999; Schafer et al., 2007; Cremonesi et al., 2008; Konstan & Riedl, 2012; Bobadilla et al., 2013; Jeckmans et al., 2013).

The first collaborative filtering algorithms were mainly user-to-user type and used variants of a weighted, k-nearest-neighbor (kNN) prediction algorithm (Konstan & Riedl, 2012). In fact, kNN approach is still the most widely used collaborative filtering algorithm when both user-to-user and item-to-item approaches are taken together (Cremonesi, Garzotto, & Turrin, 2012a; Bobadilla et al., 2013). In user-to-user type, first the most similar users are found for the current user, *u*, i.e. a *neighborhood* for *u* is formed, and then the ratings from the neighborhood for items that *u* has not rated are aggregated, and top N items are used as recommendations (Redpath et al., 2010; Konstan & Riedl, 2012; Bobadilla et al., 2013). Usually, the ratings of the neighbors are normalized and mixed into a weighted average with the similarity between the rater and *u* used as the weight (Konstan & Riedl, 2012).

Various metrics that typically have statistical origins are used to compute the similarity between users to find the *k nearest neighbors* (Bobadilla et al., 2013; Jeckmans et al., 2013). The most commonly used traditional metrics include *Pearson correlation*, *cosine*, *adjusted cosine*, *constrained correlation*, *mean squared difference*, and *Euclidean* (Bobadilla et al., 2013). The same similarity measures can also be used to calculate item-to-item similarities (Cremonesi et al., 2011; Bobadilla et al., 2013). Given their central position and importance, developing metrics for calculating similarities between users and between items is a recurring theme in collaborative filtering research (Bobadilla et al., 2013).

While the classic user-to-user collaborative filtering approach provides high-quality recommendations, the challenge for it has turned out to be its slowness when the numbers of users grow into hundreds of thousands or millions (Linden, Smith, & York, 2003; Konstan & Riedl, 2012). Item-to-item filtering, developed as an alternative to user-to-user approach, has a significantly faster online response time, especially if the item relationships are pre-computed, in addition to offering slightly better recommendations (Linden, Smith, & York, 2003; Schafer et al., 2007; Cremonesi et al., 2008; Campochiaro et al., 2009; Konstan & Riedl, 2012).

Item-to-item algorithms are the transpose of the user-to-user algorithms (Schafer et al., 2007). Item-to-item approach builds correlations between pairs of items and then generates recommendations by finding items that are similar to the set of items that the current user has (implicitly or explicitly) shown liking (Schafer, Konstan, & Riedl, 1999; Schafer et al., 2007; Konstan & Riedl, 2012). In other words, if user-to-user approach is based on *similar users liking similar items*, item-to-item approach is based on *users liking items similar to the items they have shown preference for* (Jeckmans et al., 2013). Thus, in both, the idea is to generate recommendations based on the preferences of users with similar tastes to the current user in the past, i.e. to ignore the content of the item and exploit collective preferences of the crowd (Cremonesi, Garzotto, & Turrin, 2012a). In fact, one of the great strengths of collaborative filtering is that since they are not dependent on any machine-readable aspect of the item being recommended, they work well with any kind of items, including complex objects where variation in user opinions are largely due to variations in user tastes (Burke, 2002). Also, collaborative filtering, together with demographic techniques, has the "capacity to identify cross-genre niches and can entice users to jump outside of the familiar" (Burke, 2007).

While the user-to-user approach requires at least some preference data from the user, item-to-item recommendations do not need a user profile for generating recommendations; in fact, they can be generated as soon as the user has expressed interest in just one item (Schafer, Konstan, & Riedl, 1999). For example, *Customers who…* recommendations on Amazon.com use item-to-item approach and do not require users to sign in or even to have visited the site previously to provide recommendations (Schafer, Konstan, & Riedl, 1999; Linden, Smith, & York, 2003). Being able to generate recommendations in this way is important for any business, as many returning clients of Amazon.com do not necessarily sign in when they start to browse items (Leino, 2008).

The strengths of the user-to-user and item-to-item versions of the kNN algorithm can also be combined (Bobadilla et al., 2013).

Latent factor approach is the second major approach to collaborative filtering. Latent factor models "represent users and items as vectors in a common low-dimensional 'latent factor' space" where "users and items are directly comparable and the rating of a user … on an item can be estimated as the proximity … between the related latent factor vectors." Due to its good performance in terms of root-mean-squared error metric, this approach has been very successful in the Netflix contest. (Cremonesi et al., 2011)

### Inherent Challenges in Collaborative Filtering Approach

All algorithmic approaches have their advantages and disadvantages that make them better fit to certain contexts and challenging to apply in some other contexts, and this also applies naturally to collaborative filtering (Burke, 2002; Thor, Golovin, & Rahm, 2005; Schafer et al., 2007).

One major challenge for collaborative filtering are ***scalability*** problems (Schafer, Konstan, & Riedl, 2001; Schafer et al., 2007; Su & Khoshgoftaar, 2009; Park et al., 2012). In naïve implementations, the time and memory requirements of user-to-user algorithms "scale linearly with the number of users and ratings", making it impossible to use such approaches e.g. on Amazon.com (Schafer et al., 2007; see also Su & Khoshgoftaar, 2009). Such challenges have been tackled e.g. with *subsampling*, using a subset of users selected prior to prediction computing, and *clustering*, comparing a user to a group of users instead of to individual users and then selecting the nearest neighbors from the clusters most similar to the user (Schafer et al., 2007). However, the problem has not been entirely solved, and with various implementation approaches, tradeoffs between scalability and prediction performance tend to persist (Su & Khoshgoftaar, 2009).

Another major challenge to collaborative filtering is ***data sparsity***, i.e. a sparse user-item matrix resulting from few users rating the same item (Burke, 2002; Konstan et al., 1997; Schafer et al., 2007; Su & Khoshgoftaar, 2009; Park et al., 2012). Collaborative filtering depends on overlap in ratings (or other preference data) across users, and if few users have rated the same items, correlations either cannot be calculated or can be skewed—with little overlap, neighborhood forming is challenging and correlation based on too few common ratings result in spurious correlations (Konstan et al., 1997; Burke, 2002; Schafer et al., 2007). Consequently, as far as data distribution is concerned, collaborative filtering works the best when there are many items and many ratings per each item, "[t]here are more users rating items than items to be recommended," and "[u]sers rate multiple items" (Schafer et al., 2007; see also Burke, 2002).

Data sparsity also leads to ***coverage*** problems, i.e. the percentage of items in the system that predictions can be calculated for is low (Herlocker et al., 2004; Schafer et al., 2007; Su & Khoshgoftaar, 2009). Moreover, it leads to ***cold-start*** issues, i.e. the system is unable to make meaningful recommendation because of the lack of preference data for new users, new items, and new communities (Konstan et al., 1997; Burke, 2002; Schafer et al., 2007; Su & Khoshgoftaar, 2009; Bobadilla et al., 2013). A neighborhood cannot be calculated to a *new user* as the new user's similarity with the old ones cannot be calculated as no or little preference data exists, i.e. no user model/profile exists (Burke, 2002; Su & Khoshgoftaar, 2009; Konstan & Riedl, 2012). Consequently, casual users are excluded from receiving the full benefit of the system; it is dedicated users that get the most benefit from learning-based systems (Burke, 2007). While such recommendations as *Customers who…* can be given to new users, such benefits as deep personalization that collaborative filtering at its best offers are not possible (Burke, 2002; Konstan & Riedl, 2012). A *new item* problem is basically the same: an item that has not been rated cannot be recommended (Burke, 2002; Su & Khoshgoftaar, 2009).

Another problem related to data sparsity and new items is ***first-rater*** problem—the first person to rate an item receives little benefit from rating it, as rating it does not allow calculating their neighborhood any better (Konstan et al., 1997; Burke, 2002). In a *new community* situation, the first-rater, or perhaps early-adopter, problem persists for a long time, as necessary coverage needs to be achieved before the system begins to produce value to the users (Konstan et al., 1997).

The data sparsity problem, when personalization is a must, often results in the use of hybrid recommenders where collaborative filtering is combined with e.g. content-based, demographic-based or social-based approaches (Konstan & Riedl, 2012; Bobadilla et al., 2013). However, all learning-based techniques, be they collaborative, content-based, or demographic, are subject to cold-start problems one way or another (Burke, 2007). Cold-start problems can be made less severe by carefully selecting which items are given for a user to rate, as this can have considerable effects on how quickly good recommendations can be generated to the user (Konstan & Riedl, 2012). For example, Rashid et al. (2002) studied six techniques to select items for new users to rate, concluding that the best performance came from a hybrid that used popular items that had high ratings entropy[16].

Schafer et al. (2007) also posited that for collaborative filtering to work, *for each user, there needs to be other users with common needs or tastes* and *item evaluation should require personal taste*. If there are no users who share a user's need or taste, predictions cannot be calculated for the user (Su & Khoshgoftaar, 2009; Konstan & Riedl, 2012). This is known as ***black sheep*** problem—users have too idiosyncratic tastes (Su & Khoshgoftaar, 2009). ***Gray sheep*** problem, in contrast, refers to users who do not consistently agree or disagree with any group of people, thus falling between existing cliques (Burke, 2002; Su & Khoshgoftaar, 2009). As to item evaluation requiring personal taste, if the goodness of the item can be calculated based on objective criteria, then there is little need for collaborative filtering (Konstan & Riedl, 2012).

Finally, collaborative filtering needs *items to persist* and *tastes to persist* (Konstan & Riedl, 2012). If items change quickly, it is challenging to have many users rate them to create an overlap in ratings (Konstan & Riedl, 2012). An example where the collaborative filtering approach has difficulties is news recommending, as the approach requires overlapping ratings but news stories are most interesting when they are new and fresh (Cleger-Tamayo, Fernández-Luna, & Huete, 2012; Konstan & Riedl, 2012).

[16] Entropy measures disagreement among raters (Konstan & Riedl, 2012).

As to tastes, if tastes do not persist over time but change quickly, then older ratings are not useful for collaborative filtering, and, again, generating neighborhoods is challenging (Konstan & Riedl, 2012).

Finally, collaborative filtering systems tend to be biased towards popularity—popular items tend to get recommended, resulting in a rich-get-richer effect and reduced diversity (Cremonesi, Garzotto, & Turrin, 2012a; Konstan & Riedl, 2012). Lack of serendipity and diversity can lower user satisfaction (Cremonesi et al., 2011; Cremonesi, Garzotto, & Turrin, 2012a; Knijnenburg et al., 2012; Pu, Chen, & Hu, 2012). In fact, the early emphasis on accuracy metrics to evaluate collaborative filtering systems has resulted in a mismatch between user needs and evaluation metrics; error and correlation scores do not properly assess how well an algorithm can recommend valuable items that are unknown to the user, i.e. the items that the collaborative filtering approach is designed to recommend to begin with (Konstan & Riedl, 2012).

## 5.3 Content-Based Filtering

Content-based filtering represents a continuation of information filtering research (Burke, 2002). The recommendations are based on attributes associated with items and preference data, e.g. ratings that users have given to items (Burke, 2007). While collaborative filtering is based on the assumption that people with similar tastes rate things similarly, content-based filtering is based on the assumption that items with similar objective features, or attributes, are rated similarly (Schafer et al., 2007; Bobadilla et al., 2013). In effect, content-based recommenders suggest to a user items the content of which is similar to the content of the items that the user has rated positively, or has otherwise shown preference for, in the past (Cremonesi et al., 2011; Cremonesi, Garzotto, & Turrin, 2012a; Bobadilla et al., 2013). In other words, a content-based approach learns a profile of the user's interests based on the attributes present in the items that the user has rated positively (Burke, 2002).

For example, if the user has liked a web page with words like "car", "engine", and "gasoline", the system will suggest more pages related to automobiles (Bobadilla et al., 2013) or, in case of videos, the content can be the title, the actors, the director, and the genre of the liked videos, as videos can be characterized by them and they are readily available (Cremonesi, Garzotto, & Turrin, 2012a). In other words, the item similarity is calculated based on explicit content attributes associated with the items being compared (Cremonesi, Garzotto, & Turrin, 2012a; Jeckmans et al., 2013).

One of the challenges is to identify and extract the attributes, or meta-data, of items that are most predictive (Schafer et al., 2007; Bobadilla et al., 2013).

Text domains work well for extracting keywords and such meta-data but some content, e.g. sound file content (rather than associated metadata, e.g. artist), can be hard to analyze (Cosley et al., 2003; Schafer et al., 2007). If some attribute cannot be automatically extracted from the item, it has to be added by hand or it cannot be used in content-based filtering (Schafer et al., 2007). In addition to extracting attributes, a user profile must be built to know preferred items that can then be compared to other items for attribute similarity (Schafer et al., 2007; Park et al., 2012).

While the basic approach to content-based filtering bases its similarity analysis on term-by-item occurrences, neglecting the semantic structure of the content, more advanced approaches, such as latent semantic analysis, attempt to also exploit semantic features (Cremonesi et al., 2011). In addition to keywords and other meta-data that can be extracted, content-based systems are increasingly incorporating social information on items that users in Web 2.0 provide, such as tags, posts, and reviews (Bobadilla et al., 2013).

#### Inherent Challenges in Content-Based Filtering Approach

Content-based filtering approach in its pure form has numerous shortcomings (Burke, 2007; Bobadilla et al., 2013). In certain domains, generating attributes for items is challenging but content-based systems are, by their very nature, limited to the attributes that are explicitly associated with the items they are to recommend (Burke, 2002; Bobadilla et al., 2013). For example, a content-based movie recommender is limited to written materials about a movie (Burke, 2002). Consequently, a "content filtering model can only be as complex as the content to which it has access" (Schafer et al., 2007).

In addition, because these systems suggest items the content of which is similar to the content of the items that the user has shown preference for in the past, content-based approaches also suffer from ***overspecialization*** problem (Schafer et al., 2007; Cremonesi, Garzotto, & Turrin, 2012a; Park et al., 2012; Bobadilla et al., 2013). The user is often already aware of the items that are suggested or could have figured them out easily—e.g. by searching for movies with the name of their favorite actor or books by their favorite author—and so recommendations lack unexpected quality, or serendipity (Schafer et al., 2007; Cremonesi, Garzotto, & Turrin, 2012a; Park et al., 2012; Bobadilla et al., 2013).

Also, like collaborative filtering, content-based filtering suffers from ***cold-start*** problems because they need to know some items that the user has rated positively as a starting point (Burke, 2007). Without enough ratings, content-based filtering cannot build a reliable classifier (Burke, 2002).

Because of its inherent shortcomings, content-based filtering is rarely employed in its pure form (Bobadilla et al., 2013). Often, it is combined with collaborative filtering as a hybrid to overcome problems in both (Schafer et al., 2007; Bobadilla et al., 2013). In fact, in many ways, content-based and collaborative filtering are complimentary (Schafer et al., 2007). In domains where content is scarce or difficult to obtain, collaborative filtering can function efficiently because users take care of the evaluation and little needs to be known about the item beyond this (Schafer et al., 2007). On the other hand, content-based filtering can function without ratings e.g. for new items, for high-turnover items (e.g. news stories), and huge item spaces (e.g. web pages), as items can be evaluated based on attributes, while collaborative filtering without preference data on an item is unable to consider it (Schafer et al., 2007). However, the two approaches can be combined in multiple ways, and no consensus exists on how to do it the best (Schafer et al., 2007).

## 5.4 Clustering Approaches

Clustering approaches divide the user base into segments, or groups of users who have very similar preferences, treating the task as a classification problem. Predictions for a user are then calculated based on averaging the opinions of other users in the cluster. In some clustering approaches, a user can have partial participation in numerous clusters, and the predictions are then based on the average across the clusters of participation, weighted by degree of participation. (Schafer, Konstan, & Riedl, 2001; Linden, Smith, & York, 2003)

Clustering techniques, once clustering is done, can have a good online scalability and performance, as the size of the group to be analyzed is smaller than e.g. in collaborative filtering. However, the recommendations generated are not very well personalized and the quality can be low. (Schafer, Konstan, & Riedl, 2001; Linden, Smith, & York, 2003)

Thanks to its strengths, clustering can be used as a first step method that shrinks the candidate set. While the accuracy may somewhat suffer for users who are on the fringes of clusters, the improved performance may be a worthwhile tradeoff. (Schafer, Konstan, & Riedl, 2001)

## 5.5 Demographic Filtering

Demographic information, such as gender, age, and country of residence, can be used to generate somewhat personalized recommendations (Bobadilla et al., 2013; Jeckmans et al., 2013). The idea is that users with common demographic attributes also have common preferences (Bobadilla et al., 2013). Demographic data is gathered through various means, e.g. by surveys or through machine learning (Burke, 2002). The demographic information is first matched to a stereotype and then the items connected to it are recommended (Jeckmans et al., 2013) or the ratings in a demographic niche to which the user belongs are combined to produce recommendations (Burke, 2007). Personalization is naturally limited when users are generalized in this manner (Jeckmans et al., 2013).

The good side of demographic filtering is that it does not require a history of user ratings and, consequently, does not suffer from *new user* problems (Burke, 2002). Consequently, demographic filtering has been combined e.g. with collaborative filtering to deal with cold-start issues (Bobadilla et al., 2013). Demographic data has also been suggested to be used to lessen cold-start problems outside of pure demographic filtering approaches, see e.g. Nguyen, Denos, & Berrut (2007). However, demographic recommenders also have to gather the demographic information somehow, so they are not free from data gathering challenges altogether (Burke, 2002). Finally, demographic filtering suffers from the ***gray sheep*** problem, of some users falling between the existing stereotypes (Burke, 2002).

## 5.6 KNOWLEDGE-BASED, UTILITY-BASED, AND CRITIQUING SYSTEMS

Pu, Chen, and Hu (2012) consider both utility-based and knowledge-based systems as variants of case-based systems. *Case-based* systems are based on case-based reasoning techniques that solve new problems using a case database of past problem solving experiences, retrieving a similar case and adapting its solution to the current problem. In case-based recommender systems, items are represented as cases and recommendations are generated by picking the case, i.e. items, that correspond most closely to the user query or profile. In comparison to content-based filtering, case-based systems rely on a more structured representation of item content and they use various similarity assessment approaches for identifying similar cases. (Pu, Chen, & Hu, 2012)

*Utility-based* systems generate recommendations by computing the utility of each object for the active user. Even non-product attributes, e.g. product availability, can be factored into the utility computations. However, this flexibility is also a challenge in utility-based systems, as creating a utility function for each user means that each user has to construct a complete preference function, considering the significance of every possible feature. In most cases, this constitutes a significant burden for the user, at least in the case of more complex and subjective domains. Some systems have used interactive questionnaires to derive the function. While some users, e.g. technical users with specific requirements, may cherish building a utility function, more casual users with less detailed domain knowledge are likely to be overwhelmed. Utility-based approach can be seen as a special case of knowledge-based approach. (Burke, 2002)

*Knowledge-based* systems also require the user to input their preferences (Burke, 2002; Jeckmans et al., 2013). After the user has input their preferences, the system presents them with recommendations based on the knowledge contained in the system (Jeckmans et al., 2013). After a few iterations of the process, the recommendations are tailored to the user. In learning knowledge-based systems, feedback from the user is used to add to the knowledge (Jeckmans et al., 2013).

The problem for all knowledge-based systems, including recommenders, is that there is a need for knowledge acquisition (Burke, 2002). Burke (2002) lists three types of knowledge necessary in a knowledge-based recommender system:

1) *Catalog knowledge*: Knowledge about the items to be recommended and their attributes; e.g. a restaurant recommender needs to know that "Thai" cuisine represents a type of "Asian" cuisine.
2) *Functional knowledge*: The system has to "have knowledge about how a particular item meets a particular user need" so that they can "reason about the relationship between a need and a possible recommendation"; e.g. a restaurant recommender needs to know that a restaurant that is "quiet with an ocean view" is a good candidate for a romantic dinner place (see also Burke, 2007).
3) *User knowledge*: The system also needs to have some knowledge about the user, perhaps "general demographic information or specific information about the need for which a recommendation is sought." The latter case is especially tricky, as it can, in the worst case, descend into "the general user-modeling problem."

Overall, knowledge-based systems are more suited to casual browsing than utility-based ones, as they demand less effort and knowledge from the user (Burke, 2002).

Knowledge-based and utility-based approaches do not need to accrue user profiles; instead, they "base their advice on an evaluation of the match between a user's need and the set of options available". Consequently, they do not suffer from cold-start or sparsity problems, since the recommendations are not based on accumulated statistical data. Also, there is no ***plasticity*** problem, a problem that occurs in other approaches that base recommendations on user modeling when the user preferences change, e.g. a steak-eater becomes a vegetarian. While collaborative and content-based filtering systems continue to feed the new vegetarian steakhouse recommendations for quite some time, utility-based and knowledge-based systems that do not rely on historical profile data do not need any re-training and can respond immediately to the current user need. (Burke, 2002)

In addition to utility-based and knowledge-based systems that Burke (2002, 2007) discusses, Pu, Chen and Hu (2012) also consider *critiquing-based* systems, a third type of case-based systems. Critiquing systems represent a form of case-based systems that operate in a reactive fashion, simulating a salesperson that recommends items based on the user's current preferences and then elicits feedback in form of critiques from the user, e.g. "I'd prefer a faster processor" (Pu, Chen, & Hu, 2012). The critiques from the user allow the recommender to refine its recommendations until a satisfactory item is found (Pu, Chen, & Hu, 2012). In this way, critique-based systems allow users to construct and refine their preference model incrementally (Pu, Chen, & Hu, 2012). In a sense, critique-based systems are interactive utility-based systems, as they allow constructing the utility function interactively under the guidance of the system.

## 5.7 Personality-Based Systems

Pu, Chen, and Hu (2012) consider personality "an enduring and primary factor that determines human behavior" as it is connected to a person's tastes and interests. They suggest that since personality-based recommenders understand users better in psychological sense, they can therefore be used to provide deeper personalization. Consequently, they see personality acquisition techniques as an emerging field that has employed both implicit (observing user's behavior to infer personality) and explicit (personality questionnaires) approaches. However, personality-based recommender approaches are still in the early stages of research, and it remains to be seen how influential the approach becomes. Still, it appears to offer at least additional information on which to base recommendations.

## 5.8 Context-Aware Systems

Context-aware recommender systems aim to generate more relevant suggestions by adapting them to the user's contextual situation (Adomavicius et al., 2011; Jeckmans et al., 2013). By traditionally focusing on two entities, users and items, recommender systems have ignored the notion of *situated actions*, that decision making is "contingent on the context of decision making" (Adomavicius & Tuzhilin, 2011; see also Adomavicius et al., 2011). For example, if the user is buying a book for their child or a work-related book for themselves, the context of the task is clearly different and the recommendations should also be different, even though it is the very same user for whom recommendations are generated (Adomavicius et al., 2011). In the same way, a user may like to read stock-market information in the morning in the office and movie reviews in the weekend, and the company with which a user is planning to see a movie can also affect as to what kind of movies should be recommended (Adomavicius & Tuzhilin, 2011). Some companies have started to incorporate contextual information into their systems, e.g. Sourcetone interactive radio[17] takes into consideration the mood of the listener (that the listener has to explicitly specify) (Adomavicius & Tuzhilin, 2011).

Adomavicius et al. (2011) discuss four broad types of context that were used by context-aware applications that they reviewed: *Physical context* (e.g. time, position, and temperature), *social context* (presence and role of others; is the user alone or in some group), *interaction media context* (device the user is using, e.g. mobile phone, and the type of media browsed and personalized, e.g. music, text, or queries), and *modal context* (e.g. mood, goal). Contextual information can be obtained in various ways, including *explicitly* (e.g. asking the user to fill out an online form), *implicitly* (from the data or the environment, e.g. time from the timestamp of the transaction), and through *inference* (inferring by using statistical or data mining methods, e.g. inferring which family member is watching TV from the programs being watched) (Adomavicius & Tuzhilin, 2011).

Adomavicius and Tuzhilin (2011) state that "the context-aware recommendation process that is based on contextual user preference elicitation and estimation can take one of the three forms":

1) ***Contextual pre-filtering***, or contextualization of recommendation input: Contextual information drives data selection or construction for the specific context.
2) ***Contextual post-filtering***, or contextualization of recommendation output: Contextual information is used to adjust the set of ratings that were generated without using contextual information.
3) ***Contextual modeling***, or contextualization of recommendation function: "[C]ontextual information is used directly in the modeling technique as part of rating estimation."

Context-aware recommender systems represent a relatively new and unexplored field (Adomavicius & Tuzhilin, 2011). They represent one of the directions that recommender research is likely to take in the future (Adomavicius et al., 2011; Konstan & Riedl, 2012; Bobadilla et al., 2013).

## 5.9 Hybrid and Ensemble Recommenders

The fact that all algorithms used in recommender systems have their complementary strengths and weaknesses has led to many kinds of combinations to combat their weaknesses, e.g. cold-start issues (Burke, 2002, 2007; Xiao & Benbasat, 2007; Bobadilla et al., 2013). Also, besides using other algorithmic approaches to cover weaknesses in one, approaches are combined to be able to combine their strengths, e.g. to take

[17] http://www.sourcetone.com/

advantage of both the item attributes that the user likes (content-based) and community preferences (collaborative) (Xiao & Benbasat, 2007).

A *hybrid* recommender systems is "any recommender system that combines multiple recommendation techniques together to produce its output" (Burke, 2007; see also Jeckmans et al., 2013). While several different techniques of the same type can also be combined to improve performance, they are not considered hybrid recommender systems (Burke, 2007; Jeckmans et al., 2013); instead, systems that combine several of the same type of techniques are known as *ensemble* recommenders (Jeckmans et al., 2013).

#### Ensemble Recommenders

The term *ensemble* is typically "reserved for collections of models that are minor variants of the same basic model" (Rokach, 2009). As a method, the ensemble approach imitates the way we often ask several people for their opinion before making an important decision, weighing them individually and combining them for the final decision (Rokach, 2009). The idea of an ensemble recommender is to combine a set of models that all solve the same task in order to obtain a better global model than any of the single models used in the combination would provide alone (Schclar et al., 2009). According to Schclar et al. (2009), experiments have shown the superiority of the ensemble approach over any one model. They themselves used an ensemble approach to improve the predictive performance of neighborhood-based collaborative filtering. For a classification of ensembles, see Rokach (2009).

#### Hybrid Recommenders

While hybrid recommender studies have already proven that the approach is beneficial and has produced many successes, no hybrid has, naturally enough, emerged as a be-all-end-all of recommender techniques (Burke, 2007). Consequently, more work is necessary 1) to understand what kinds of hybrids are likely to succeed and what kinds are unlikely to do so and why, and 2) to determine in which domains and with what data characteristics various hybrids work the best (Burke, 2007). The most comprehensive overview and referred to paper on hybrid recommender systems is by Burke (2007) (see e.g. Jeckmans et al., 2013; Bobadilla et al., 2013), and the reader is consequently referred to that paper for further detail on hybrid systems.

## 5.10 Social Recommenders

In a sense, even the early versions of user-to-user collaborative filtering were implicitly based on social ties, namely similarity between users, which allowed the formation of neighborhoods. In effect, such recommender systems use implicit social networks present in their user base that can be induced from explicit and implicit preference data to generate recommendations (Lerman, 2007). Today, some recommendation sites, e.g. Digg.com, "allow users to explicitly construct social networks and use them for getting personalized recommendations" (Lerman, 2007). The rise of social networks has, in fact, greatly increased the availability of social information, e.g. friendship networks, followers and followed, and tags, so recommender systems are increasingly incorporating this information (Bobadilla et al., 2013; Jeckmans et al., 2013). The underlying assumptions in this trend are that since friends typically share interests, 1) the information that a user's friend provides fits the user more likely than the information provided by somebody whose similarity is based on statistical inference, and 2) that communities of similar users can be inferred from social information (Jeckmans et al., 2013). For example, Konstas, Stathopoulos, and Jose (2009) used explicitly expressed friendship information to improve the recommendations on Last.fm in comparison to the standard user-to-user collaborative filtering. However, while Bobadilla et al. (2013) assert that "users connected by a network of trust exhibit significantly higher similarity on items and meta-data than non-connected users", Doerr et al. (2012) found that at least on Digg.com, this does not necessary hold, as discussed below.

Broadly speaking, social information has been used for three primary objectives, 1) to improve the quality of recommendations, 2) to propose or build new recommender systems, and 3) to "elucidate the most significant relationships between social information and collaborative processes" (Bobadilla et al., 2013).

While there are many positive examples of using social information to augment or generate recommendations, e.g. Konstas, Stathopoulos, and Jose (2009) and Sen, Vig, and Riedl (2009), there are also indications that social networks may not always be as important as perceived (Doerr et al., 2012). While it is generally thought that "the friendship relations within a social network are the critical ingredient to the successful spread of information", Doerr et al. (2012) found that "the importance of these links and their contribution is very low".

There is, in fact, on Digg.com—the social network where Doerr et al. (2012) made an analysis of ten million stories and two million users—only a probability of 2% that users "with even a nearly identical overlap in interests" react to "information propagated and received from friends." Also, in about 50% of the stories that managed to make it to the front page, "there was no prior contribution by the friend network"—the critical mass came from random spectators. In the other cases, a story was predominantly pushed by friends of the submitter even though non-friends then needed to make significant contributions for the story to reach the critical mass. Based on their analysis, Doerr et al. (2012)

concluded that "the vast majority of similarly behaving user pairs" had not formed a "friendship" between them. In effect, while there were "some patterns in the common behavior of users, the bare topology of the friendship graph is unsuited to fully capture it." (Doerr et al., 2012)

Of course, the domain of Digg.com—that of fast-moving news stories—affects the equation and the social information on Digg.com differs from many other social networks. Consequently, the results of Doerr et al. (2012) should not be generalized too far. Nevertheless, their study does underline two facts, 1) recommenders work in various very different domains, so over-generalizing case study results needs to avoided, and 2) social network information is not always the best way to approach things, no matter how popular the approach may appear today. That said, it does seem evident that the use of social data in the realm of recommender systems is set to increase and will result in high quality recommendations. In a sense, in fact, Facebook.com is a huge social recommender system.

## 5.11 DELIVERING AND PRESENTING RECOMMENDATIONS

In terms of the taxonomy by Schafer, Konstan, and Riedl (2001), presenting the output of recommender systems is classified as part of *functional I/O*, specifically functional output (*input* is concerned with preference elicitation). *Output* is concerned with how recommendations vary in terms of type, quality, and look of information that is provided to the user (Schafer, Konstan, & Riedl, 2001).

### Delivering Recommendations

Before delving deeper into the presentation issues, however, let us look briefly at *delivery* aspects, i.e. how the output ends up before the eyes of the user. *Push* technologies reach out to the user the way direct mail and telemarketing does in the traditional setting. The most common push technology for delivering recommendations to the user is email. The goal is to invite the user back to the site employing the recommender. *Pull* technologies put the user in control of when recommendations are displayed. The user is made aware of recommendations being available, e.g. by displaying a link to them, but they are displayed only after the user asks for them, e.g. by clicking the link. *Passive* delivery, also known as *organic* delivery, "presents the recommendation in the natural context of the rest of the" site. For example, Amazon.com makes an extensive use of passive delivery in form of e.g. *Customers who...* and *Frequently bought together…* recommendations that are displayed on the product page in the context of an item that the user has selected for viewing. Consequently, passive recommendations can reach the user at the point where they are receptive to the idea. (Schafer, Konstan, & Riedl, 2001)

Many recommenders provide a prediction of the rating that the user would give to the items. These predictions may be in personalized form, i.e. how the active user would rate each item, or in non-personal form, e.g. average of star ratings given to each item, thus providing a community estimate. Editorial ratings have also been employed. Predictions are given both in the context of the suggestion lists and on the page where the item information is displayed. (Schafer, Konstan, & Riedl, 2001)

Ratings are also useful in the context of reviews, as reviews are evaluations that are not completely machine-understandable; with textual recommendations, it is hard to know whether the recommendation is *for* or *against* the item without reading and interpreting it. Ratings allow users to see this with a glance and the systems to know it. Text recommendations, although typically non-personalized, allow users to see why a certain rating has been given and help users to evaluate items. Sometimes, e.g. on Amazon.com, reviews can also be meta-rated, and these meta-ratings can be used to order and select reviews. (Schafer, Konstan, & Riedl, 2001)

#### Importance of the Interface and Presentation

The interface of the site and of the recommendation system may significantly influence the user experience (Knijnenburg & Willemsen, 2009). In general, users "subconsciously process information embedded in the design of a technology" and "use perceived control and effort as important judgment cues" (Gretzel & Fesenmaier, 2006). In recommender systems, while content-related factors have a more significant role in determining user satisfaction than structural factors, interface factors, e.g. navigation and layout, do influence the equation strongly when they present obstacles to the user (Swearingen & Sinha 2001; Xiao & Benbasat, 2007; Ozok, Fan, & Norcio, 2010). Consequently, features and information have to be easily accessible, meaning that navigational paths and a clear layout constitute important design aspects of the recommender systems (Swearingen & Sinha 2001; Xiao & Benbasat, 2007).

Moreover, recommender systems as decision aids can influence the user perceptions of the quality of the recommended items, as the way recommendations and related information are shown in the interface affects the user evaluations of the recommendations (Gretzel & Fesenmaier, 2006). The way recommendations are formatted can also strongly influence the user's decision process and the outcome of that process (Xiao & Benbasat, 2007; Knijnenburg et al., 2012).

#### Recommendation List Related Factors

The most common type of algorithmic recommender output is a *suggestion* of an item that typically takes the form of "try this" or simply placing "this" in the page that the user is viewing. While a single item represents the simplest form of this approach, it is more common that a set of suggestions is provided in a particular context. Some lists are apparently unordered (although of course there is some ordering involved) while others are given as ranked, or ordered, lists that often provide extra information to the user. (Schafer, Konstan, & Riedl, 2001)

Given the importance of the presentation and the fact that most recommendations are displayed as lists, it is surprising that the composition of the recommendation list has received as little research attention as it has (Knijnenburg et al., 2012). Pu, Chen, and Hu (2012) use the term *set composition* to refer to the choice in the list length (how many items to suggest) and how various kinds of items are mixed together and balanced to form the final set. It is still unclear how many recommendations should be included on the recommendation list displayed to the user, and serial positioning effects remain unclear even though we know that the order in which the recommendations are presented does matter (Knijnenburg et al., 2012).

The display method for presenting recommendations affects the evaluation strategies that users use during the initial, or screening, stage of the decision-making process. As the typical goal of an algorithmic recommendation system is to recommend a list of salient items, many items on the list are often very similar in the overall subjective utility to the user but may differ on several attributes. When the differences are large, it is easy to decide, but if the differences are small, the user must make many attribute comparisons between any two items to reach an informed decision. However, as the choice difficulty increases, users tend to start using simplifying heuristics instead of using utility-based comparisons of evaluating the current item against the most attractive item they have found thus far. (Xiao & Benbasat, 2007)

Consequently, Xiao and Benbasat (2007) suggest that users "may rely more on heuristic decision strategies when distinguishing among the product alternatives" on a sorted list of personalized recommendations (i.e. sorted based on how well they are predicted to match user preferences) and in deciding which ones to include in the consideration set. In fact, sorted lists of personalized recommendations can result in users having higher decision quality and paying lower prices in comparison to randomly ordered lists (Xiao & Benbasat, 2007).

One aspect of the suggestion list that has received some research attention is the optimal number of items on the list. In fact, if the set of high-quality recommendations is too large, it can result in *choice overload* (Bollen et al., 2010), resulting in reduced selectivity and poor item choices, i.e. decreased decision quality, and increased effort in evaluating alternatives, i.e. increased decision effort (Xiao & Benbasat, 2007). In this way, recommendation systems may actually reduce information overload without making the decision stage any easier (Bollen et al., 2010).

In effect, although users are attracted to larger item sets, they in fact find it easier to choose from smaller sets and are generally happier with their choices, as choosing from a large set of good items can be challenging. In fact, attractiveness of a larger set of high quality items does not necessarily result in higher user satisfaction, as the increased difficulty of choosing an item offsets the initial attraction. (Bollen et al., 2010)

However, not all large item sets appear to result in choice overload (Bollen et al., 2010). It seems that a large set results in choice overload when users have no strong prior preferences on which to ground their choices, when there are no dominant options among the items, and when there is little difference in the attractiveness of the items (Bollen et al., 2010). In effect, a large set of high-quality items increases cognitive cost, i.e. makes choosing more difficult, while a large set of items containing inferior items does not (Bollen et al., 2010; Knijnenburg et al. 2012).

Consequently, Bollen et al. (2010) wonder if larger sets are necessary when a small set can already incorporate enough salient items without causing choice overload. While they emphasize that further research is necessary, they suggest that "seven to ten items would be very attractive, quite varied, and still be manageable for the users." Pu, Chen, and Hu (2012) suggest that five items might be the upper limit, stating that "[s]tudies indicate that while showing one item is too few, showing more than five items increases users' choice difficulty, but increases their perception of diversity."[18] Ozok, Fan, and Norcio (2010) go even lower, stating that, based on a survey study, users want a maximum of three recommendations.

On the other hand, there are also some reasons to increase the number of items in the recommendation set. First, a larger set appears more attractive to the user (Bollen et al., 2010). Second, a larger set allows for more *diversity* and increases perceived diversity that positively affects user satisfaction (Ziegler et al., 2005; Konstan, 2008; Bollen et al., 2010; Hu & Pu, 2011; Pu, Chen, & Hu, 2012). Finally, a larger set allows for *familiarity*, i.e. including familiar items on the list to increase trust in the recommender system, without having to use familiar items to replace novel recommendations (Sinha & Swearingen, 2002; Herlocker et al., 2004; Pu, Chen, & Hu, 2012). In effect, accuracy is clearly not the only concern in the user satisfaction with the recommended items (Ziegler et al., 2005; Bollen et al., 2010; Hu & Pu, 2011). It may even be that the presence of inferior items on the list provides a decision context that reduces choice difficulty (Pu, Chen, & Hu, 2012).

[18] However, Pu, Chen, and Hu (2012) do not explicate which studies lead them to this conclusion in their paper.

In effect, the implication is that individually good recommendations put together do not automatically equal a good recommendation list (Konstan, 2008). Users are likely to treat and experience recommended items as a set or as a whole rather than seeing them as individual items—the item(s) that the user has viewed influence how they see the subsequent items (McNee, Riedl, & Konstan, 2006b; Pu, Chen, & Hu, 2012). Consequently, the "recommendation list should be judged for its usefulness as a complete entity," not "as a collection of individual items"; lists are more than aggregations of individual recommendations and have inherent, added value (McNee, Riedl, & Konstan, 2006b; see also Ziegler et al., 2005). Importantly, users are aware of qualitative differences in lists, including the breadth or depth of the list vis-à-vis their goals (McNee, Riedl, & Konstan, 2006b). Finally, the user task and context also affect how users react to the recommendation lists (Redpath et al., 2010).

Awareness of these factors has resulted in attention shifting to a larger extent to the recommendation lists and judging them as wholes instead of considering only the quality of each individual recommendation (Konstan & Riedl, 2012).

#### What Information to Show

While it is necessary to provide enough information on the recommended items for the recommendation to facilitate the decision-making process (Pu, Chen, & Hu, 2012), how much is enough remains a somewhat open question. According to Xiao and Benbasat (2007), available empirical evidence points out that detailed product information positively affects users' evaluations of recommender systems: Detailed information about the items increases trust on, perceived usefulness of, and satisfaction with the recommender systems.

An online survey study by Ozok, Fan, and Norcio (2010) of 131 college students found that shoppers like to see precise product information. They found that the essential information is the product name, price, and image, and that shoppers additionally value peer feedback in the form of customer ratings and peer comments (secondary type of information). According to them, no further details are generally wanted on the suggested products.

In a study of music recommender system interfaces involving 12 users, Swearingen and Sinha (2002) found that users use several types of information in making up their minds. *Basic item information* included song, album, and artist name; genre; and year of release, in addition to a picture of the album cover. The second type of information was *Expert and community ratings*; reviews and ratings by other users seemed especially important. Moreover, *Item sample* was appreciated; in the domain of music, it is rather natural that users liked to evaluate the recommendation by listening to a sample of the song (Swearingen & Sinha, 2002).

In a study of three book and three movie recommender systems involving twenty users, Swearingen and Sinha (2002) found that longer descriptions correlate positively with perceived usefulness and ease of use of the system. In contrast, Ozok, Fan, and Norcio (2010) found that shoppers like short, concise descriptions of recommended items. It seems that at least some differences in the two studies are due to different domains and perhaps, to some degree, to different methodologies. For example, the wish for a sample is natural in a music setting while not even possible in some others. Also, in case of, say, hotels, it might be that the basic information would include the location of the hotel in the city. Domain is likely to strongly affect what is considered basic or essential information.

That said, besides basic information that is necessary to judge the item at all (e.g. item name), both studies emphasized the importance of community opinion in the form of ratings and reviews. eWOM clearly matters to users in making up their minds about the recommended items.

#### Diversity

Often, the items on the recommendation list are very similar to each other; e.g. books by the same author (Ziegler et al., 2005; McNee, Riedl, & Konstan, 2006a). Item-to-item collaborative filtering seems especially prone to producing this kind of recommendations, i.e. recommendations suffering from *portfolio effect* (Ziegler et al., 2005; McNee, Riedl, & Konstan, 2006a). Still, if judged by accuracy metrics, such recommendations can be extremely good (Ziegler et al., 2005). The problem is more noticeable the less data there is to base the recommendations on, e.g. in the case of a new user (McNee, Riedl, & Konstan, 2006a). Recommendations should not only be accurate but also *useful*, but this is something that accuracy metrics cannot capture (Herlocker et al., 2004; McNee, Riedl, & Konstan, 2006a).

User satisfaction is increased by *diversity* in the recommended list, and, consequently, diversity should be a goal even at the cost of lower accuracy, as increased diversity affects user satisfaction more than reduced accuracy (Ziegler et al., 2005; McNee, Riedl, & Konstan, 2006a; Pu, Chen, & Hu, 2012). However, it is important to note that the user perception of diversity does not appear to correlate with the algorithmic understanding of diversity, as humans appear able to perceive only a certain level of diversification in a list—users do not necessarily recognize more diversified lists as being more diverse (Ziegler et al., 2005; Knijnenburg et al., 2012; Pu, Chen, & Hu, 2012).

Ziegler et al. (2005) approached the problem of similarity by introducing *intra-list similarity metric* and a process to diversify recommendation lists. Their results underline that we need to move beyond pure accuracy metrics to improve user experience and satisfaction, as also noted earlier by Herlocker et al. (2004). Diversity is one metric that appears connected to user satisfaction even if accuracy is reduced (Ziegler et al., 2005).

### Familiarity

*Familiarity* refers to there being some items with which the user is familiar on the list of recommended items (Swearingen & Sinha, 2002; Pu, Chen, & Hu, 2012). In fact, while novelty is considered more useful than familiarity, users like familiar recommendations and having them present in a set of recommendations increases trust in the recommendation system and the recommendations it generates (Swearingen & Sinha, 2002; Xiao & Benbasat, 2007; Pu, Chen, & Hu, 2012). This is in spite of the fact that users are well aware that recommendations of novel items are more useful (Xiao & Benbasat, 2007).

Naturally enough, however, too many familiar recommendations or recommendations too directly related to the input ratings, e.g. albums by an artist whose album the user has rated positively, is frustrating to users, as such recommendations do not help users expand their horizons (Swearingen & Sinha, 2002). Conversely, seeing only unfamiliar recommendations also lowers evaluations (Xiao & Benbasat, 2007), likely because judging the quality of the recommendations becomes harder. Consequently, a good set of recommendations is a suitable mixture of novel or serendipitous recommendations and familiar items, as novel recommendations are the goal of the user but familiarity is needed to establish trust (Swearingen & Sinha, 2002; Herlocker et al., 2004; Xiao & Benbasat, 2007). However, in establishing trust it is naturally important to get predicting right, i.e. to recommend familiar items that users actually like, as recommending familiar items they do not like would naturally lower their trust in the system (Swearingen & Sinha, 2002).

Herlocker et al. (2004) see familiar recommendations especially important for the users whose task is to *find credible recommender*. Consequently, Konstan and Riedl (2012) suggest that for new users, a recommender system should provide enough familiar items to establish trust, while the emphasis should be more on novelty for users who are already familiar with the system. In effect, there needs to be a balance between novelty, or usefulness, and familiarity, or trust-building, which may vary during the recommender use life-circle.

### Novelty and Serendipity

The opposite of familiarity is the "non-obviousness" of a recommended item (Herlocker et al., 2004). After all, highly accurate recommendations can be practically useless (Herlocker et al., 2004; McNee, Riedl, & Konstan, 2006a). For example, recommending *The White Album* to a person who has expressed a preference for The Beatles is highly accurate but practically worthless; they either have the album or have decided not to have it but they are, in both cases, highly likely to be aware of the album (Herlocker et al., 2004). In comparison, a recommendation of a garage band that a fan of The Beatles would like but that they would never learn about through reviews or advertisements would be a much more valuable recommendation (Herlocker et al., 2004). In effect, while familiar recommendations do have their value, as discussed above, for many applications, non-obviousness, being directed to items we did not know about is one of the most valued characteristics of a recommender system (Herlocker et al., 2004; Schafer et al., 2007). For example, in terms of sales effect, recommendation novelty has a stronger effect than recommendation accuracy (Cremonesi, Garzotto, & Turrin, 2012a).

*Novelty*, a dimension of non-obviousness that has been discussed in information retrieval literature (Herlocker et al., 2004), refers to the ability of a recommender system to suggest items that the user was not aware of (Schafer et al., 2007). *Serendipity*, on the other hand, refers to the ability of a recommender system to recommend surprising items that the user would not have otherwise discovered (Herlocker et al., 2004; Schafer et al., 2007). In effect, serendipitous recommendations are always novel but novel recommendations are not always serendipitous (Herlocker et al., 2004). As an example of the difference, Herlocker et al. (2004) discuss a user who is not aware of a movie by their favorite director. As a recommendation, that movie would be novel but not really serendipitous, as the user would likely have discovered the movie on their own, too, e.g. by a simple search. The distinction between novelty and serendipity is important[19], as serendipity is something that collaborative filtering systems can provide while content-based filtering cannot, even though both can provide novelty (Herlocker et al., 2004).

While accuracy is easy to measure with statistical measures, serendipity is not, as it involves "the experience of receiving an unexpected and fortuitous item recommendation" (McNee, Riedl, & Konstan, 2006a), of being recommended an item that is both attractive and surprising (Herlocker et al., 2004a). Measuring even novelty is hard, as it really requires live user studies so that the participants can indicate whether or not the recommended item was novel (Schafer et al., 2007). In fact, accuracy metrics have a tendency to reward algorithms that make the most obvious recommendations (Herlocker et al., 2004). There is a certain tradeoff to be considered between high quality as defined by accuracy and as defined by novelty (Herlocker et al., 2004).

Today, serendipity as a concept has been extended to a wide range of control over recommendations, including diversity and similarity (Konstan & Riedl, 2012).

---

[19] Nevertheless, in their ResQue framework for user-centric evaluation of recommender systems, Pu, Chen, and Hu (2011) saw novelty as being "related to the recommender's ability to educate users and help them discover new items" and dropped the concept of serendipity, as "the meticulous distinction between these two words will cause confusion for users."

### Transparency: Explaining Recommendations to the User

One reason why recommender systems have been more successful in low-risk domains (e.g. books, CDs, and movies) than in high-risk ones (e.g. cars and real-estate) appears to be that users are not willing to risk that much based on recommendations that they do not understand (Herlocker, Konstan, & Riedl, 2000; Konstan & Riedl, 2012). In fact, recommender systems tend to be black boxes that provide little transparency into the workings of how recommendations are generated (Herlocker, Konstan, & Riedl, 2000; Schafer et al., 2007). Users have been left with little ground to decide if they should trust the recommendations (Schafer et al., 2007).

Explanations can provide *transparency* as they can explain how the recommendation was formed, thereby allowing the user to determine how far they wish to trust it and therefore giving them confidence to act on it (Konstan & Riedl, 2012). While recommendations suggest items to users in which they are likely to be interested, explanations explicate *why* they might be interested in them (Vig, Sen, & Riedl, 2009). Moreover, users like explanations, explanations influence how users perceive the recommender system, and explanations help users make better decisions (Herlocker, Konstan, & Riedl, 2000; Vig, Sen, & Riedl, 2009; Gedikli, Ge, & Jannach, 2011). Consequently, Knijnenburg et al. (2012) consider explaining why an item was suggested to the user an important element of the recommendation presentation. In fact, Ozok, Fan, and Norcio (2010) go as far as to state that “transparency is the criterion that primarily should be used to evaluate the usefulness of the interface of the recommender systems.”

Since the first study of providing explanations for recommendations by Herlocker, Konstan, and Riedl (2000), inspired by the considerable benefit that providing explanations in expert systems provides, study after study (e.g. Sinha & Swearingen, 2002; Vig, Sen, & Riedl, 2009; Ozok, Fan, & Norcio, 2010) has shown that providing transparency through explanations is vital. Importantly, explanations and transparency increase trust in the recommender system in general—a decision system that its users do not trust is, after all, not a very useful one—and also on its competence and benevolence (after all, e-retailers may have strong motivations to manipulate their recommenders for their own benefit) (Herlocker, Konstan, & Riedl, 2000; Gretzel & Fesenmaier, 2006; Xiao & Benbasat, 2007; Pommeranz, Wiggers, & Jonker, 2011; Knijnenburg et al., 2012; Pathak et al., 2010). Moreover, transparency increases confidence in the recommendations (Herlocker, Konstan, & Riedl, 2000; Gretzel & Fesenmaier, 2006; Schafer et al., 2007; Knijnenburg et al., 2012; Bobadilla et al., 2013) and transparency is an important factor that influences whether a recommendation is accepted and positively evaluated (Gretzel & Fesenmaier, 2006). In fact, Schafer et al. (2007) consider a direct relationship to exist between the user being explained the recommendation and the user being persuaded by the recommendation.

Moreover, explanations can better involve users in the recommendation process, allowing them to add their knowledge and inference skills to the decision-making process, and they can educate users as to the recommendation process so that they better appreciate the strengths and limitations of the system (Herlocker, Konstan, & Riedl, 2000). In effect, transparency and explanations can increase user satisfaction and help users make better decisions (Xiao & Benbasat, 2007; Vig, Sen, & Riedl, 2009). Another factor in the equation is that while user trust in the system appears to influence their experience of it, their experience of system also affects their trust in the system (Knijnenburg et al., 2012).

Overall, transparency plays a highly influential role in the success of the recommenders (Ozok, Fan, & Norcio, 2010). Given the importance of transparency and explanations, it is no wonder that features such as Amazon.com's *Why was I recommended this item* have started to pop up in other systems, too, e.g. Netflix and Pandora (Schafer et al., 2007; Pu, Chen, & Hu, 2012).

Interestingly, it appears that gathering user preferences explicitly results in more confidence in the recommendations (Gretzel & Fesenmaier, 2006; Cremonesi, Garzotto, & Turrin, 2012b) while implicit preference data collecting makes generated suggestions harder to understand (Pommeranz, Wiggers, & Jonker, 2011). It appears that explicit preference eliciting provides some transparency that helps the user to understand what the system does to provide recommendations (Pommeranz, Wiggers, & Jonker, 2011).

Providing transparency through explanations has proven "challenging because the underlying predictive models are complex aggregations of large quantities of data, often with significant probabilistic reasoning" (Schafer et al., 2007). At the same time, simple explanations are more effective than sophisticated explanations of how algorithms work, as users can easily be overwhelmed by too involved explanations (Herlocker, Konstan, & Riedl, 2000; Schafer et al., 2007; Konstan & Riedl, 2012).

A common approach to providing explanations for recommendations—why a certain item was recommended to the user—is to use an *intermediary entity* to explain the relationship between the user and the item (Vig, Sen, & Riedl, 20099). Intermediary entities are typically *items*, *users*, or *features* (Figure 2) (Vig, Sen, & Riedl, 2009). Consequently, explanations typically fall into one of the four categories, *item-based*, *user-*

*based*, *feature-based*, or *hybrid methods* (Vig, Sen, & Riedl, 2009; Bobadilla et al., 2013)[20].

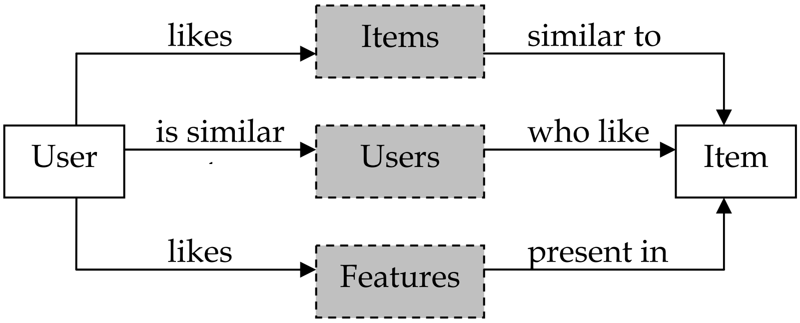


**Figure 2**. How intermediary items relate users to recommended items (Vig, Sen, & Riedl, 2009).

*Item-based* explanations use *items* as intermediary entities to connect the user to the recommended item (Vig, Sen, & Riedl, 2009); e.g. we recommend the book *i* because you liked books *j*, *k*, *m*… (Bobadilla et al., 2013). Item-based explanations improve the user acceptance of the recommendations, in addition to helping users make accurate decisions, but a shortcoming with them is that the user may fail to grasp the relation between the explaining items and the recommended item (Vig, Sen, & Riedl, 2009).

*User-based* explanations use other *users* as intermediary entities: "The relationship between the current user and the explaining users is typically that they share similar tastes, and the relationship between the explaining users and the recommended item is that they have rated the item positively" (Vig, Sen, & Riedl, 2009); e.g. we recommend the book *i* because it was liked by users who also liked books *j*, *k*, *m*… (*j*, *k*, *m*… being books that the current user has also rated positively) (Bobadilla et al., 2013). User-based explanations have persuasive power but are less effective at helping users in making accurate decisions (Herlocker, Konstan, & Riedl, 2000; Vig, Sen, & Riedl, 2009).

*Feature-based* explanations use qualities or characteristics, e.g. *predefined categories* and *keywords* have been used, of the recommended items as intermediary entities (Vig, Sen, & Riedl, 2009); e.g. we recommend the book *i* because it was written by *j* and belongs to the genre *k* (*j* and *k* being characteristics that interest the current user) (Bobadilla et al., 2013).

*Hybrid methods* category is mainly composed of *human/item*, *human/feature*, *feature/item*, and *human/feature/item* explanations (Bobadilla, et al., 2013).

[20] Vig, Sen, and Riedl (2009) only discuss three types of explanations, *item-based*, *user-based*, and *feature-based*, while Bobadilla et al. (2013) adds the *hybrid method* as the fourth. Incidentally, Bobadilla et al. (2013) call the three other types by slightly different names than Vig, Sen, and Riedl (2009): *Item style explanations*, *Human style explanations*, and *Feature style explanations*.

With their *tagsplanation* approach of using a tag or a set of tags as the intermediary entity, Vig, Sen, and Riedl (2009) showed that explanations can use a completely different dimension than algorithms use and still be effective (Konstan & Riedl, 2012). In effect, with the *tagsplanation* approach, the goal was *justification* rather than *transparency*; while transparency aims at users understanding what the system is doing (Pommeranz, Wiggers, & Jonker, 2011), justification might not reveal the actual mechanism of the recommender algorithm while still successfully helping the user to understand why a certain item was recommended (Vig, Sen, & Riedl, 2009).

Gedikli, Ge, and Jannach (2011) study of using tags in an explanation interface in form of tag clouds[21] showed that explanations can also be personalized and that personalized explanations can provide value over non-personalized approaches.

[21] For further detail on the study, see Section 2.6 (discussion on Explaining Recommendations with Tags).

# 6 eWOM: Community Opinion

As discussed in Section 2.5, in a sense, all recommender systems are from the start modeled after informal word-of-mouth (WOM), i.e. designed to assist in the natural social process of people using the opinions and experience of other people, in particular the people they know, to help in decision making. In effect, recommender systems allow us to go beyond the restraints of traditional word-of-mouth by allowing us to consider and use the opinions of thousands of people instead of tens or hundreds when we need to make choices under uncertainty (Hill et al., 1995; Schafer et al., 2007). While algorithmic recommenders typically provide recommendations or predictions, eWOM provides community opinion.

While eWOM has largely been studied in the context of e-commerce and consequently defined in commerce-related terms, we argue that using it e.g. in the context of study materials is not out of place, as students do exchange opinions about study materials (and teachers, for that matter) in WOM (e.g. face-to-face and over the telephone) and eWOM (e.g. social networking sites) settings. Consequently, we argue that the eWOM concept can be applied outside of the e-commerce context, too, and, consequently, can be equated with the idea of *community opinion*[22].

## 6.1 Defining WOM and eWOM

Hu, Pavlou and Zhang (2006) define traditional WOM as "all informal communications" that consumers "direct at other consumers about ownership, usage, or characteristics of particular goods and services or

[22] Schafer, Konstan, and Riedl (2001) do not define eWOM as community opinion—they do not even use the term eWOM—but as eWOM is defined, it quite closely matches their definition of community opinion; see Section 2.5.

their sellers" while Cheung and Lee (2012) define it as "an oral form of interpersonal noncommercial communication among acquaintances". Cheung and Lee's (2012) definition limits WOM to take place between acquaintances, and Hu, Pavlou, and Zhang (2006) also see WOM as being limited to "local social network". WOM is widely seen as a major factor in how new products fare among non-adopters (Hu, Pavlou, & Zhang, 2006) and as significantly influencing how products are evaluated, as well as which products end up getting purchased (Cui, Lui, & Guo, 2012).

The advent of the Internet has brought on a paradigm shift in WOM; WOM has evolved into eWOM, electronic word-of-mouth, that takes place in various online settings: Users can post opinions, comments, and reviews of items on blogs, discussion forums, reviews sites (e.g. Epinions.com), e-retail websites (e.g. Amazon.com), e-bulletin boards, newsgroups, and social networking sites (e.g. Facebook.com) (Hennig-Thurau et al., 2004; Pathak et al., 2010; Cheung & Lee, 2012). Again, the "technologists" (Schafer, Konstan, & Riedl, 2001) may object when a site such as Epinions.com, where users can rate and review products, is referred to as a recommender system, but the ratings and reviews there are undeniably "made available as *recommendations* to others" (Schafer et al., 2007, emphasis added).

Hennig-Thurau et al. (2004) define eWOM as "*any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet.*" In a similar vein, Mudambi and Schuff (2010) define eWOM as "*peer-generated product evaluations posted on company or third party websites.*" As we extend eWOM to domains beyond e-commerce, we define eWOM as *user-generated item evaluations posted online*, thus removing the emphasis on peers and customers and extending it to apply beyond products and companies. For example, Ratemyprofessors.com[23] is a service that allows students to rate professors they have studied with on Overall quality, Helpfulness, Clarity, Easiness, and Hotness certainly represents eWOM. It is clearly outside the domain of e-commerce, is provided by students rather than consumers (even if some would claim that students are, in fact, consumers of educational goods), and concerns not products or companies but educators in higher education.

eWOM differs from traditional WOM in several regards: eWOM communication is facilitated by technology and entails multi-way asynchronous exchanges of information (Cheung & Lee, 2012); the scalability and the speed of diffusion of eWOM is much greater than that of WOM (Hu, Pavlou, & Zhang, 2006; Cheung & Lee, 2012; Cui, Lui, & Guo, 2012); eWOM is much more persistent and accessible than WOM (Cheung & Lee, 2012; Cui, Lui, & Guo, 2012); and eWOM is no longer tied

[23] http://www.ratemyprofessors.com/

to a local community, as any consumer who has Internet access can view eWOM content from anywhere in the world, meaning that socially and geographically dispersed strangers can easily share opinions and experiences related to items (Hu, Pavlou, & Zhang, 2006; Cheung & Lee, 2012). By being persistent, available if not indefinitely then typically at least for a prolonged period of time, eWOM communication is much more observable and measurable than traditional WOM (Cheung & Lee, 2012). Since eWOM decouples WOM from the local community, from one's acquaintances, it may also reduce the receiver's ability to judge the message based on its sender on such factors as credibility (Cheung & Lee, 2012). On the other hand, eWOM is convenient, e.g. easily available when the user wants it, and it is free of social pressures present in face-to-face exchanges (Cui, Lui, & Guo, 2012). All this adds up to make eWOM much more influential than WOM ever was (Cui, Lui, & Guo, 2012).

Also, unlike WOM, because of its persistence, eWOM can be viewed as *public good*, or a shared resource that can benefit every member of a group, independent of whether they have contributed to it or not, and the availability of which does not diminish from consumption (Cheung & Lee, 2012). Similarly, eWOM, as well as algorithmic recommender systems, can be seen as *value-added services* in that they reduce search cost and uncertainty involved in item choosing (Pathak et al., 2010). Still, at the same time, persistence can also have a negative side: If the item reviewed changes, the old eWOM no longer represents the reality of the item but continues to affect consumer perceptions. For example, a hotel that has been reviewed poorly may get new owners who renovate it thoroughly but the old eWOM continues to describe it as seedy, driving consumers away (Ludford, 2007).

eWOM literature typically focuses on product reviews that consist of a numerical rating, often on a five-star scale, and open-ended comments (Pathak et al., 2010; Mudambi & Schuff, 2010) even though many authors do point out that eWOM is not limited to them; e.g. Pathak et al. (2010) speak of "digital word-of-mouth, where shoppers can submit and share their feedback about products through reviews and rating systems on retailer Web sites" only to later speak of "various forms of digital word of mouth, such as online product ratings and reviews, social network sites, blogs, and so forth." As discussed, we also include tags into eWOM, as tags individually and as aggregates (e.g. tag-clouds) provide further information about the tagged item (Lee, Son, & Han, 2007), e.g. by pointing out a category to which the item belongs, and can also provide opinions about items (Wal, 2007), functioning as kind of mini-reviews (Leino, 2012b).

In practice, still, eWOM literature tends to focus on item (or product) evaluations that consist of a rating and a review/comment. Reviews are typically accompanied by ratings, as free textual comments are slow to read and the strength of the recommendation for or against may not be clear to the reader; the rating makes it clearer (Schafer, Konstan, & Riedl, 2001; Talwar, Jurca, & Faltings, 2007). In effect, star rating has been found to function as a cue for the review content for users (Mudambi & Schuff, 2010). On the other hand, the review can give justification for the rating in addition to providing further information (Schafer, Konstan, & Riedl, 2001). Also, as machines cannot make complete sense of textual reviews (Schafer, Konstan, & Riedl, 2001) and ratings allow reviews to be e.g. grouped by the ratings as Amazon.com does, different users have different strategies for finding relevant reviews, e.g. some like to read negative reviews (Leino, 2008).

According to Cui, Lui, and Guo (2012), three metrics of consumer product reviews have been subjected to closer examination: Volume, valence, and dispersion. *Volume* is important because the higher the volume, the higher the increase in awareness among consumers; e.g. the volume of messages on a new movie is a good predictor of the box office success. *Valence*, typically the average of ratings or the fraction of positive and negative opinions, provides community opinion about the item quality, thereby providing either a positive or negative recommendation. The importance of valence is evident in how eWOM studies have successfully used ratings to forecast revenue for new products; e.g. the valence of online ratings during a movie's opening week is an important predictor of the revenue it will garner. *Dispersion*, or spread, refers to how fast eWOM spreads within and across communities; e.g. it has been shown that eWOM can affect product sales very early in the process. (Cui, Lui, & Guo, 2012)

## 6.2 Importance of eWOM

The mere presence of eWOM on a website improves consumer perception of the usefulness and social presence of the site, as does the presence of algorithmic recommenders, too (Kumar & Benbasat, 2006). For the e-retailers, they offer potential to create a sense of community among the users, resulting in increased *stickiness* (time spent on site) and number of visits (Mudambi & Schuff, 2010).

In turn, users use eWOM for making better, more informed decisions, for reducing uncertainty about the item quality and suitability, and for lowering transaction costs (Hu, Liu, & Zhang, 2008; Mudambi & Schuff, 2010). Significantly, by enabling consumers to make better decisions, eWOM also leads to consumers experiencing greater satisfaction (Mudambi & Schuff, 2010). Moreover, information seeking can be in and of itself a source of pleasure for at least some consumers (Svensson, Höök, & Cöster, 2005; Mudambi & Schuff, 2010).

Users often have to make purchasing or other decisions with incomplete information. As seeking information is costly, e.g. requiring effort and time, there is an inevitable trade-off between perceived costs and benefits of acquiring additional information, and the information seeking costs increase the total costs, adding to the product costs. At the same time, the user knows that, while additional information can reduce uncertainty, uncertainty cannot be eliminated entirely, so there is an inevitable trade-off between effort and accuracy, too. (Mudambi & Schuff, 2010)

eWOM can effectively help users to reduce the amount of effort and time needed to get additional information and can ease and improve the purchase decision process (Mudambi & Schuff, 2010). Consequently, eWOM has quickly grown to be one of the most important sources of information for consumers (Hu, Pavlou, & Zhang, 2006; Racherla & Friske, 2012), enabling them to judge the quality of various products, services, and other types of commercial interactions (Talwar, Jurca, & Faltings, 2007). In effect, Cui, Lui, and Guo (2012) state that "online product review Web sites outrank all other media in influencing customer decisions" and that the "user-generated content, especially online product reviews, helps consumers make informed decisions about purchasing new products and has become a major driving force in new product sales, making effective e-marketing a critical success factor for new product launch." It has been claimed that, as a result of the eWOM explosion online, consumers are today more informed about what is available than ever before (Racherla & Friske, 2012).

Importantly, eWOM has been found to be more potent than printed information because eWOM is seen as more credible and valuable (Cui, Lui, & Guo, 2012). In effect, eWOM can make or break products, especially new products for the sales of which it has become a major driver (Cui, Lui, & Guo, 2012). Consequently, eWOM "has very important implications for a wide range of management activities, such as brand building, customer acquisition and retention, product development, and quality assurance" (Hu, Pavlou, & Zhang, 2006).

In 2011, according to Cone Online Influence Trend Tracker, 89% of the U.S. consumers found the Internet a trustworthy source for product and service reviews, and 85% went online to do additional research before deciding whether or not to purchase something they had been recommended. Moreover, 80% of consumers had changed their mind about a purchase based solely on the negative information they had read online, up from 69% in 2010. Positive online information had also a significant impact, as 87% of consumers agreed that a positive review had confirmed their purchase decision. The upward trend in comparison to 2010 can be attributed to the increasing ubiquitous Internet access and pervasiveness of smartphones. (Cone Online Influence Trend Tracker, 2011)

In effect, eWOM has today significant influence on consumer behavior and any manager that ignores it does so at their own peril (Cone Online Influence Trend Tracker, 2011; Cheung & Lee, 2012). The impact of online reviews is such that some consumers are even willing to pay 20% more for services receiving a five-star rating than the same service receiving a four-star rating (Talwar, Jurca, & Faltings, 2007; Cheung & Lee, 2012).

eWOM is especially important when the nature of the product under consideration makes it difficult to evaluate (Pathak et al., 2010). In the case of services, eWOM is particularly important as they are intangible by nature (Racherla & Friske, 2012).

However, while eWOM provides a wealth of important information for users, at the same time it may in and of itself be a source of information overload, as items can and do have hundreds, even thousands of reviews; e.g. Khaled Hosseini's *And the Mountains Echoed* had 5,472 customer reviews on Amazon.com[24] on April 1, 2014. In addition, research has identified many biasing factors in how eWOM is today collected, aggregated, and listed, and companies can and do disguise their promotions as consumer-based recommendations. (Racherla & Friske, 2012)

Despite its increasing importance, eWOM has only recently started to attract growing academic research interest. Today, it is a topic of great interest in various fields, especially business-oriented disciplines such as consumer behavior and economics, but also in various fields related to information systems. The main research interest focuses on the impact of eWOM. (Cheung & Lee, 2012)

## 6.3 Factors That Affect eWOM Impact

Today, the significance of eWOM is beyond question. The question now for many e-retailers and services in other domains is how to provide high quality eWOM rather than whether or not to provide it. In fact, as eliciting eWOM can be cumbersome, Amazon.com is generating additional revenue by selling eWOM to other players. High-quality reviews and providing easy access to them can, in fact, become a differentiation factor for a site. (Mudambi & Schuff, 2010)

Mudambi and Schuff (2010) define a helpful customer review as a "*peer-generated product evaluation that facilitates the consumer's purchase decision process.*" They see *helpfulness* "as a reflection of review diagnosticity" in

[24] http://www.amazon.com/Mountains-Echoed-Khaled-Hosseini/dp/159463176X/

that eWOM can provide *diagnostic value* across the various phases of the purchase decision process; once a need has been recognized, eWOM assists consumers in both information search and evaluating alternatives, in addition to some consumers also using it in the post-purchase evaluation phase. In fact, consumers' assessment of the customer review usefulness has been shown to be an important antecedent to information adoption both online and offline (Racherla & Friske, 2012).

Investigating the influence of eWOM on users centers on the question of "*who* says *what* to *whom* with *what* effect". The *who* refers to the characteristics of the sender, or message source, including such factors as status and expertise; the *what* refers to the characteristics of the message, including such factors as quality, length, and valence; and the *whom* part refers to the receiver characteristics, including expertise, prior knowledge, and level of involvement. (Racherla & Friske, 2012)

Users—the *whom* of the above equation—do not constitute a uniform group but diverge into subgroups along various dimensions that have been studied in the field of consumer psychology. One such dimension is *need for closure* (NFC), or the desire for unambiguous and definite knowledge to guide the consumer perception and action. In studying NFC in a low-involvement purchase situation, Vermeir, Van Kenhove, and Hendrickx (2002) found that when confronting a new choice situation, people with high NFC search for a lot of information at once to be able to make a confident decision, as they experience the absence of closure as uncomfortable. They also use numerous attributes in comparing products. They see the amount of information used in making a decision as correlating with the decision quality. They are typically quite certain about their decisions. In contrast, people with low NFC only seek and use as much information as is necessary for them to feel that they have enough knowledge to make a satisfactory choice. This amount is not large, as they do not consider it necessary to spend a lot of energy on a low-involvement decision. In effect, both people with high NFC and with low NFC experience a lot of uncertainty to begin with in a new choice situation but the amount information they seek and use for making the choice is very different. (Vermeir, Van Kenhove, & Hendrickx, 2002)

Source-related factors are seen as having informational and normative influences (Chen, 2008; Racherla & Friske, 2012). *Informational influence* refers to accepting the information as evidence about reality and is thought to operate through *internalization* that occurs if the reference group is considered credible (Chen, 2008; Racherla & Friske, 2012). *Normative influence*, on the other hand, refers to consumers conforming to and complying with the positive expectations of others (Chen, 2008; Racherla & Friske, 2012). There are two types of normative influence, utilitarian and value-expressive influence: *Utilitarian influence* refers to the situations where individuals conform to the expectations of others to gain rewards or to avoid punishment, while *value-expressive influence* refers to the situations where influence operates through *identification*, whereby individuals wish to associate with groups that they evaluate positively or to distance themselves from groups that they evaluate negatively, thus maintaining or enhancing their self-concept (Racherla & Friske, 2012).

The influence of a review is affected by the number of reviews available; the higher the number of existing reviews, the less weight a new review adds, because once a critical mass of reviews has been reached, new reviews typically can only add a limited amount of new information[25] (Hu, Liu, & Zhang, 2008). Also, consumers only expect a few of the reviews to be useful for them and reliable, and so it is not really the sum of all reviews that influences the purchase decision but the few reviews that the consumer judges relevant (Leino & Räihä, 2007; Racherla & Friske, 2012). Also, when reading reviews, consumers deal with two types of uncertainty, with one type dealing with uncertainty about the item of interest, e.g. its quality and characteristics, and uncertainty about the integrity and intentions of the reviewers (Racherla & Friske, 2012). Consequently, consumers use both social and informational cues in assessing the reliability of a review (Racherla & Friske, 2012).

#### Item Type

The type of the product or service—search vs. experience vs. credence—influences what kind of a review and what kinds of review characteristics constitute a useful review (Mudambi & Schuff, 2010; Yang & Mai, 2010; Cui, Lui, & Guo, 2012; Racherla & Friske, 2012). *Search items*, be they products or services, are typically evaluated based on relatively objective attributes, e.g. technical specifications, that can be evaluated before purchase, and therefore represent a low level of risk (Mudambi & Schuff, 2010; Yang & Mai, 2010; Cui, Lui, & Guo, 2012; Racherla & Friske, 2012). *Experience items*, on the other hand, cannot be evaluated without experiencing them, resulting in increased uncertainty and greater reliance on eWOM as a risk-coping strategy (Mudambi & Schuff, 2010; Yang & Mai, 2010; Cui, Lui, & Guo, 2012; Knijnenburg et al., 2012; Racherla & Friske, 2012). In contrast, *Credence items*, e.g. a haircut, cannot be evaluated even after consumption (Yang & Mai, 2010; Racherla & Friske, 2012). However, typically items are divided into search and experience items[26] (Mudambi

[25] Of course, the context and domain can affect this, as a recent review of a hotel can be considered as adding valuable information even if there are already numerous older reviews.

[26] Interestingly, even though books have been used as an item in many eWOM studies, there is apparently no consensus on whether books are search or experience items; Knijnenburg, Willemsen, and Kobsa (2011) and Chen (2008) consider them experience goods while Xiao and Benbasat (2007) consider them search goods. In effect, given that books have different attributes, some of which are search attributes (e.g. a customer who does not want to read a huge book may be interested in shorter books or a reader who wants a recent book to make sure that it contains the most recent knowledge) and some experience attributes (e.g. how well a customer will like a

& Schuff, 2010). Still, as many items represent a mix of search and experience attributes, they can be seen as existing on "a continuum from pure search goods to pure experience goods" (Mudambi & Schuff, 2010; see also Yang & Mai, 2010).

The product type influences both the search behavior and the use of information sources (Mudambi & Schuff, 2010; Cui, Lui, & Guo, 2012). As search items can be judged based on more objective attributes, the decision-making process is more likely a systematic evaluation of specific product attributes. Also, it is relatively easy to compare search products based on such attributes (Mudambi & Schuff, 2010). In contrast, as experience goods are evaluated based on more subjective attributes that depend on personal taste, consumers are forced to rely more on attribute-irrelevant cues, such as popularity based on e.g. sales or community size (Mudambi & Schuff, 2010; Yang & Mai, 2010; Cui, Lui, & Guo, 2012; Racherla & Friske, 2012). In fact, Yang and Mai (2010) suggest that while online reviews cannot fully turn experience attributes into search attributes, popularity indicators, such as the size of the user base for games, can function as search attributes, effectively conveying the current users' evaluation on experience attributes.

Relatively speaking, finding actionable information on search items and evaluating them based on that information is easy, while evaluating experience items is difficult and obtaining relevant information is costly (Mudambi & Schuff, 2010). Also, consumers have different information requirements for search and information items and, consequently, the item type also affects what constitutes a helpful, useful review (Mudambi & Schuff, 2010; Racherla & Friske, 2012).

Consequently, for experience goods, the social presence that eWOM provides can be important (Kumar & Benbasat, 2006; Mudambi & Schuff, 2010). In fact, users tend to compare themselves to others and sometimes select for closer reading reviews written by reviewers who appear similar to themselves in social background, level of experience, and preferences (Leino & Räihä, 2007; Mudambi & Schuff, 2010; Racherla & Friske, 2012).

Moreover, the reviewer rating loses at least some of its significance for experience services in comparison to search services, as experience is, after all, a question of personal taste (Racherla & Friske, 2012). In effect, reviews may function more as sources of information on items than as evaluations of items for consumers (Leino & Räihä, 2007). In any case, consumers do pay attention to the textual portions of reviews, and some authors even claim that they pay more attention to the content than the statistical summaries of ratings (Racherla & Friske, 2012). Consequently, some

---

certain detective book), we see that, while dividing items into search and experience item brings important insight into the dynamics of the equation, the division is not without its problems.

reviews have a stronger impact on consumer behavior than the others (Leino & Räihä, 2007; Racherla & Friske, 2012).

#### Volume, Valence, and Variance

Cui, Lui, and Guo (2012) suggest a framework for how eWOM affects the sales of new products that gives *valence* (average of ratings) and *volume* (number of review views) as key metrics. In general, unsurprisingly, the overall effect of valence is that negative valence affects sales negatively and positive valence positively (e.g. Chevalier & Mayzlin, 2006; Hu, Liu, & Zhang, 2008; Berger, Sorensen, & Rasmussen, 2010; Mudambi & Schuff, 2010; Cui, Lui, & Guo, 2012; Racherla & Friske, 2012; Sun, 2012): When the average of ratings is high, the perceived quality goes over a certain threshold, leading to general interest, and when the average is low, most are not interested in the product (Sun, 2012).

In case of positive valence affecting sales positively, consumers receive affirmative evidence for an item that they are already considering, resulting in *confirmation bias* (Racherla & Friske, 2012). Naturally enough, consumers tend to ignore positive reviews of items that they do not intend to buy in the first place (Racherla & Friske, 2012), and so positive reviews have more of a confirming function. In the case of negative valence influencing consumer behavior, we see the impact of *negativity bias* (Cui, Lui, & Guo, 2012; Racherla & Friske, 2012). Negative eWOM can result in users not purchasing items they had intended to purchase (Cone Online Influence Trend Tracker, 2011).

In general, consumers find extreme reviews—rated at extremes of the scale in use—as more useful than middle-of-the-road ones (Racherla & Friske, 2012) although the effect appears stronger for experience goods than for search goods (Mudambi & Schuff, 2010). Negative reviews appear to have a significantly more pronounced effect than positive ones, perhaps due to the proposition of prospect theory that suggests that we experience losses more acutely than gains, resulting in risk aversion (Kahneman & Tversky, 1979; Chevalier & Mayzlin, 2006; Lee, Park, & Han, 2008; Berger, Sorensen, & Rasmussen, 2010; Yang & Mai, 2010; Racherla & Friske, 2012). Consumers also perceive negative reviews as more useful than positive or neutral reviews (Racherla & Friske, 2012). Furthermore, negative eWOM is also thought to spread faster than positive (Cui, Lui, & Guo, 2012). In effect, even if the negative reviews represent a small portion of otherwise positive reviews, they can have a significant impact (Cui, Lui, & Guo, 2012; Racherla & Friske, 2012).

However, the behavioral impact of negative reviews depends on the involvement of the user, as the quality of the negative reviews matters to high-involvement users that tend to conform to high-quality reviews, while low-involvement users tend to conform regardless of the quality of the reviews (Lee, Park, & Han, 2008). Also, brand can affect the equation, as consumers can ignore negative eWOM if they like the brand (Racherla & Friske, 2012).

Interestingly, however, there are also situations where negative valence is not detrimental to sales of an item (Berger, Sorensen, & Rasmussen, 2010; Sun, 2012). In fact, under certain conditions, it may affect sales positively (Berger, Sorensen, & Rasmussen, 2010; Sun, 2012).

Studying the effect of *The New York Times* book reviews on sales of books, Berger, Sorensen, and Rasmussen (2010) showed that negative publicity can actually increase purchase likelihood for books when there was a low prior awareness of the book, while it generally hurts the sales of books by well-known authors. In effect, negative publicity can work like an advertisement in that it can make people aware of the item and increase the item accessibility in the minds of the consumers, i.e. affect which options the consumer consider (Berger, Sorensen, & Rasmussen, 2010). In fact, Michael Jackson actually sold more albums when he was in the news for such less-than-lovable activities as being charged for child molestation and dangling his own child over a balcony (Berger, Sorensen, & Rasmussen, 2010). The explanation is that the valence of the reviews easily becomes disassociated from the item, especially if the awareness of the item was initially low, as there are few cognitive structures in the mind for unknown products to which to associate the negative valence (Berger, Sorensen, & Rasmussen, 2010). Consequently, there is no bad publicity for little known items, as negative reviews increase both purchase likelihood and actual sales (Berger, Sorensen, & Rasmussen, 2010). However, this may not apply to e.g. Amazon.com's reviews, as the reviews are on the same page with the item, meaning that there is no time lag for disassociating the negativity from the publicity (Berger, Sorensen, & Rasmussen, 2010) while in case of *The New York Times* reviews, there is such lag, as there may also be e.g. in the case of review sites that do not sell the items being reviewed, e.g. Epinions.com.

In addition, Sun (2012) showed that a higher variance in ratings, if and only if the average rating is low, leads to a higher demand and sales. In effect, a higher standard variation boosts a book's sales on Amazon.com if the average rating is lower than approximately 4.1-stars. A low average combined with high variance indicates a book that has niche-market potential, so well-matched consumers will like it. Sun (2012) sees this as an opportunity for the seller "to skim the market by selling to the best matched consumers at a premium price."

The number of reviews, or volume, for an item is used as a heuristic for assessing the general quality of the item (Yang & Mai, 2010; Racherla & Friske, 2012), and, consequently, the volume of eWOM "correlates significantly with consumer behavior and market outcome" (Cui, Lui, & Guo, 2012). Also, sheer volume can affect sales simply by increasing awareness (Cui, Lui, & Guo, 2012). However, it appears that the impact of volume is moderated by the item type: While valence has a greater impact on the sales of search products, the volume has a greater impact on experience products (Cui, Lui, & Guo, 2012), as volume can function as a popularity indicator (Yang & Mai, 2010; Cui, Lui, & Guo, 2012).

Volume appears to have a greater positive impact in the early part of the product life cycle; the effect tends to level off over time (Cui, Lu, & Guo, 2012). Also, volume affects the equation in the sense that the fewer the evaluations, the higher the impact of the existing reviews (Hu, Liu, & Zhang, 2008).

#### Other Reviewer Characteristics

***Review length***. Interestingly, we have conflicting evidence on the impact of the length of the review on its helpfulness. Mudambi and Schuff (2010) conclude that longer reviews are generally more helpful as they convey more information and, therefore, increase the consumer's confidence in the decision, although the effect is stronger for search goods than experience goods. In contrast, Racherla and Friske (2012) conclude that the word count of the review has a negligible impact on how useful it is perceived, suggesting that consumers favor reviews that are "short, sweet and to the point," given the amount of reviews available. Chevalier and Mayzlin (2006) also concluded that, based on an oft-quoted study on eWOM on Amazon.com and Barnesandnoble.com, longer reviews do not automatically translate into higher sales.

***Experts vs. peers***. Under conditions of uncertainty, people are considered to be more influenced by information from prominent and authoritative sources (Racherla & Friske, 2012). Somewhat surprisingly, when it comes to online reviews, consumers prefer consumer-generated rather than expert-generated recommendations (Chen, 2008; Racherla & Friske, 2012). In the study by Racherla and Friske (2012), expertise correlated negatively with perceived usefulness of reviews. In effect, it appears that consumers prefer guidance from those they perceive to be similar to themselves (Chen, 2008). Social similarity leads to a greater trust in reviews than expertise status, as peers have a higher perceived source credibility (Racherla & Friske, 2012). The effect is stronger for experience goods than search goods, although it exists for both (Racherla & Friske, 2012).

***Reviewer characteristics***. In general, reviewer characteristics correlate with the perceived usefulness of reviews. For example, reviewer reputation and reviewer exposure play a clear role; people are more responsive to reviews from reviewers who have a better reputation and more exposure. (Hu, Liu, & Zhang, 2008; Racherla & Friske, 2012)

#### Presenting Review Usefulness

Quality of reviews as measured by helpful ratings, see e.g. Amazon.com, can also influence item sales positively (Mudambi & Schuff, 2010). The helpful votes, or meta-ratings, for reviews are, so far, the most common approach to both present the usefulness of a particular review to consumers and to assess the usefulness and impact of reviews (Schafer, Konstan, & Riedl, 2001; Mudambi & Schuff, 2010; Racherla & Friske, 2012). For users, such meta-ratings offer ways to select which reviews to read (Leino & Räihä, 2007).

## 6.4 MANIPULATING EWOM

Some aspects of trust in eWOM have to do with trust in the party providing the recommendations and on the processes by which recommendations are generated (Xiao & Benbasat, 2007; Konstan, 2008). As many recommender systems are embedded on websites, the credibility of the website and the company or organization behind the website functions as an indicator of source credibility (Xiao & Benbasat, 2007; Konstan, 2008). Xiao and Benbasat (2007) call this construct *provider credibility*. Many providers have economic interests in inducing certain behavior in users. In effect, from the start, e-businesses employing recommenders have adapted recommenders not only to provide an enhanced customer experience but also to fit with their sales and marketing efforts e.g. by not recommending out-of-stock items or loss leaders, by optimizing inventory turn-over (e.g. Netflix), and by having human intervention in their recommender systems (e.g. Wal-Mart) (Pathak et al., 2010; Konstan & Riedl, 2012).

While at least some of these manipulations can be considered relatively innocuous—e.g. it is rather easy to claim that not being recommended out-of-stock items is good for both the e-retailer and customer—recommender systems do provide a possibility of functioning as kind of double agents that appear to help the consumer decision making but, in fact, direct them to consume according to the business needs of the service provider (Xiao & Benbasat, 2007; Pathak et al., 2010).

For example, Amazon.com used the information it had collected on its users for price targeting, i.e. for charging higher prices from people whose behavior had shown them to be price-insensitive and lower prices from those whose behavior had shown them to be price-sensitive; in effect, the past interactions determined both what the customer was offered and at what price. After Amazon's strategy was revealed, the public outcry forced it to promise not to resort to this kind of price targeting ever again. (Harford, 2007)

Amazon.com also offers a sponsored pairing program where authors and publishers can pair their books with well-selling books for a monthly fee (Pathak et al., 2010). It is very difficult for users to distinguish such recommendations from normal pair recommendations, which makes them misleading to consumers (Pathak et al., 2010). Additionally, there is evidence that Amazon.com prunes the reviews its products get by removing reviews with low ratings (Sun, 2012)—keeping in mind that due to the negativity bias, negative reviews are more effective sales killers than positive reviews are deal closers (Cui, Lui, & Guo, 2012; Racherla & Friske, 2012). Moreover, Amazon.com uses its recommenders to squeeze promotional fees from its suppliers (Packer, 2014). The fact that promotional fees affect the search results that consumers see on Amazon.com is something that most of its customers are not aware[27] (Packer, 2014).

The point is not to criticize Amazon.com's business practices but to point out that pernicious manipulations do happen. Such abuses of collected preference data as price targeting underline the serious privacy concerns involved in collecting vast amounts of preference and behavioral data implicitly and explicitly (Resnick & Varian, 1997; Lam, Frankowski, & Riedl, 2006; Kobsa, 2007; Anand & Mobasher, 2007; Schafer et al., 2007: Konstan, 2008; Jeckmans et al., 2013). Xiao and Benbasat (2007) warn retailers against pernicious manipulations, as trust and reputations are destroyed easier than established.

Of course, the service providers are not the only source that can manipulate recommender systems. Users also can and do intentionally lie, providing both behavioral data and explicit preference data to distort the recommendations that the system provides (Cosley et al., 2003; Schafer et al., 2007; Jurca & Faltings, 2009; Adomavicius et al., 2013). As the ratings in the system have an anchoring effect on other users, they can result in further distortions by influencing subsequent honest ratings (Cosley et al., 2003; Lam & Riedl, 2004; Zhang, 2011; Adomavicius et al., 2013).

In 2004, Amazon.com's Canadian site inadvertently revealed the identities of thousands of the previously anonymous review authors, revealing that book authors were writing glowingly positive reviews for their own books and flamingly negative ones for competing books[28] (Harmon, 2004). In fact, authors corral friends and family to provide positive reviews for their own books and negative ones for their competitors (Harmon, 2004). If this is not enough, there are also professional companies who specialize in selling glowing reviews to aspiring authors (Streitfeld, 2012). Liu Bing, a

---

[27] In fact, some publishers who have resisted paying "promotional fees" have seen the "Buy" button disappear from beside their books until they have agreed to pay (Packer, 2014).
[28] Then again, with literary giants such as Walt Whitman and Anthony Burgess both reviewing their own books under assumed names, the authors hawking their books on Amazon.com are at least in good literary company (Harmon, 2004).

data-mining expert at the University of Illinois, who has studied Amazon.com reviews, estimates that "about one-third of all consumer reviews on the Internet are fake" (Streitfeld, 2012).

Amazon.com also provides an example of influencing recommendations that others see by manipulating the systems with implicit ratings: A few motivated users on Amazon.com managed to get the book called *The Ultimate Guide to Anal Sex for Men* to appear in the automatically generated recommendations for a book by televangelist Pat Robertson "by repeatedly viewing the two items in sequence" (Lam, Frankowski, & Riedl, 2006).

According to a YouGov Omnibus survey (January, 2014), over a fifth of Americans (21%) who have reviewed products or services online have written reviews for products that they have not bought or even tried[29] (Diaz, 2014). As to why, the survey found that 32% had written a fake review because they "just felt like it", 22% because of not liking the idea of the product, and 23% because of not liking the manufacturer or service. What's good for individual authors is good for the big boys, too: In 2001, Sony admitted using fake quotes from critics on movie posters to promote its films (BBC, 2001).

In effect, at the face of such onslaught of manipulation, it is no wonder that different reputation systems are becoming a part of recommender interfaces, e.g. Amazon.com's *Real Name*™ badge to show that the person is writing with their own name.

[29] Then again, some of that activity has had hilarious results; see e.g. customer reviews on Amazon.com for *BIC Cristal For Her Ball Pen*: http://www.amazon.com/BIC-Cristal-1-0mm-Black-MSLP16-Blk/dp/B004F9QBE6.

# 7 Eliciting Preference Data

Functional input, or the data flowing into a recommender system, ranges from user preference data and attribute data to unstructured textual annotations, such as tags, comments, and reviews (Resnick & Varian, 1997; Schafer, Konstan, & Riedl, 2001). Schafer, Konstan, and Riedl (2001) divide such incoming data into two large categories: Targeted user inputs[30] and community inputs.

***Community inputs*** include a wide range of data concerning how numerous individuals in the community or the community as a whole see items. *Item attribute* assignments reflect community opinion by having users assign community-based labels or categories to items, e.g. film or book genre or category. *External item popularity* refers to using popularity measures from outside the recommender site as input; e.g. global box-office for movies and national bestseller lists. For example, in creating recommendation lists manually, editors can take into account this type of information in addition to the site sales figures. *Community use history*[31] (e.g. purchasing, viewing, downloading, or listening history) is implicit data that can be used e.g. to generate site-specific TopN lists or to draw conclusions about item similarities. (Schafer, Konstan, & Riedl, 2001)

While the above-mentioned community inputs are connected to the community as a whole, *text comments*, e.g. item reviews, are more directly associated with individual users (Schafer, Konstan, & Riedl, 2001). Tags would also fall into this category (see e.g. Lee, Son, & Han, 2007; Bobadilla

[30] Schafer, Konstan, and Riedl (2001) speak of targeted *customer* input rather than *user* input as they focused on e-commerce; we here enlarge the category to other domains by replacing the word *customer* with *user*.

[31] Here, too, we have replaced the e-commerce domain word *purchase* with the term *use* to enlarge this category to other domains outside of e-commerce.

et al., 2013) although Schafer, Konstan, and Riedl (2001) do not specifically mention them, since the concept of tagging did not yet exist. While text comments are helpful and well received (Schafer, Konstan, & Riedl, 2001; Swearingen & Sinha, 2002; Ozok, Fan, & Norcio, 2010), they are manual in the sense that users need to read them and interpret their meaning in general (to what degree the recommendation is positive or negative) and specifically for themselves (what the review means to themselves; e.g. is the user similarly positioned to the item as the reviewer) (Lam, Frankowski, & Riedl, 2006; Leino & Räihä, 2007; Ozok, Fan, & Norcio, 2010). Consequently, reviews are often coupled with numerical scores or ratings to simplify evaluating them (Schafer, Konstan, & Riedl, 2001). Also, the ratings of all customers provide data for producing algorithmic recommendations (Schafer, Konstan, & Riedl, 2001).

***Targeted user inputs*** are the fodder for generating personalized recommendations based on a user's current activity, long-term preferences, or both; without such data only non-personalized recommendations are possible (Schafer, Konstan, & Riedl, 2001; Neumann, 2007; Pommeranz, Wiggers, & Jonker, 2011; Cremonesi, Garzotto, & Turrin, 2012a). Targeted user inputs and community inputs as categories overlap somewhat; e.g. user ratings can be part of community input as parts of reviews but they can also be simultaneously used as targeted user input to build a user model/profile.

However, before we look at targeted user inputs in more detail, let us briefly look at terminology. Today, it is more common to talk about *preference elicitation* (e.g. Gretzel & Fesenmaier, 2006; Castagnos, Jones, & Pu, 2009; Pommeranz, Wiggers, & Jonker, 2011; Buder & Schwind, 2012; Cremonesi, Garzotto, & Turrin, 2012b; Knijnenburg et al., 2012; Pommeranz et al., 2012; Pu, Chen, & Hu, 2012) rather than *user inputs* (Schafer, Konstan, & Riedl, 1999; Schafer, Konstan, & Riedl, 2001). However, it is important to note that by *preference elicitation* many authors typically mean getting to know user preference data, the "likes and dislikes" of the user in order to create "an accurate user model needed to create system responses … adapted to the user" (Pommeranz, Wiggers, & Jonker, 2011). In other words, textual inputs are implicitly sidelined, as the purpose of recommender systems is defined to be providing personalized recommendations. In effect, we again see the definition of what recommender systems are raising its head. In this sense, *eliciting preferences* refers more to what Schafer, Konstan, and Riedl (2001) call *targeted user inputs* and leaves at least partially outside of its scope what Schafer, Konstan, and Riedl (2001) call *community inputs*. While we talk about eliciting preference data, we specifically extend it to also cover community inputs.

## 7.1 TARGETED USER INPUTS: IMPLICIT VERSUS EXPLICIT

As discussed, user preference data is collected either explicitly or implicitly. *Explicit* inputs are "intentionally made by the customer with the purpose of informing the recommender application of his or her preferences" (Schafer, Konstan, & Riedl, 2001), allowing the user "to construct and reflect upon preferences" (Pommeranz et al., 2012). Perhaps the most typical example of explicit input is a user rating an item that they have consumed (i.e. purchased, downloaded, or listened, and so forth) (Schafer, Konstan, & Riedl, 2001) although there are many ways to elicit explicit feedback, e.g. conversational methods, example critiquing, and tweaking approaches (Konstan & Riedl, 2012; Pommeranz, Wiggers, & Jonker, 2011; Pommeranz et al., 2012; Pu, Chen, & Hu, 2012). While users more typically give explicit feedback about items (Schafer, Konstan, & Riedl, 2001), other approaches have also been studied. As mentioned, Gedikli and Jannach (2010) proposed an approach where *users rate items by rating tags* so that users were, in fact, able to rate items across multiple criteria or dimensions.

*Implicit* inputs are typically inferred from user behavior without the user constructing preferences consciously and without the user necessarily being aware that data is being collected or that it is used in recommendation processes (Schafer, Konstan, & Riedl, 2001; Gretzel & Fesenmaier, 2006; Pommeranz et al., 2012). For example, a customer's purchase history or items that they are currently viewing may be used as implicit preference data in addition to dwell times and other similar measures (Schafer, Konstan, & Riedl, 2001; Buder & Schwind, 2012). Implicit data needs to be evaluated for meaning; user interests and preferences need to be extracted automatically from their actions, which is not necessarily straightforward (Gadanho & Lhuillier, 2007). At the same time, however, implicit data can be used very effectively for generating recommendations, as e.g. Amazon's *Customers who bought/viewed…* recommendations attest. Using implicit gathering of preferences closely resembles an economic theory of *revealed preferences*, according to which people reveal their preferences with the choices they make as consumers: What we do reveals our subjective values (Harford, 2007).

The approach and the interface used to eliciting user preferences influence the perceived quality of the recommender system and the recommendations it generates, impacting user decision accuracy and the intention to return (Gretzel & Fesenmaier, 2006; Knijnenburg & Willemsen, 2009; Cremonesi, Garzotto, & Turrin, 2012b): Different elicitation methods "have a substantial impact on the user experience" (Knijnenburg et al., 2012) and even "subtle variations on a given method significantly influence the user experience of" the system (Knijnenburg & Willemsen, 2009). Gretzel and Fesenmaier (2006) go as far as to state that "the preference-elicitation process is an intrinsic factor that ultimately determines a user's evaluation of the recommendation provided." As the preference elicitation is the first stage in the user-recommender interaction, it creates expectations and colors the subsequent interaction and the recommendations (Gretzel & Fesenmaier, 2006). In fact, there are numerous factors involved in preference elicitation, some of which are discussed below in more detail, that result in eliciting preferences from users not being a simple task (Pommeranz, Wiggers, & Jonker, 2011).

In spite of requiring more user effort (effort is discussed below in more detail), explicit approaches to collecting preference data are typically seen as providing more reliable data, as it does not involve extracting preferences from actions, and as providing transparency into the recommendation process that results in a slightly higher perceived recommendation quality and more confidence in the recommendations (Farzan & Brusilovsky, 2006; Gretzel & Fesenmaier, 2006; Gadanho & Lhuillier, 2007; Pommeranz, Wiggers, & Jonker, 2011; Buder & Schwind, 2012; Cremonesi, Garzotto, & Turrin, 2012b). The user feels more in control, resulting in higher user satisfaction and trust in the system (Buder & Schwind, 2012). In contrast, implicit gathering of preferences may lead to "psychological resistance" as users may react negatively to the resulting restriction of autonomy (Buder & Schwind, 2012).

Interestingly, it has also been argued that implicit preference data might in actuality be more objective, as there is no bias arising from users responding in a socially desirable way (Buder & Schwind, 2012); there are no self-image issues (how we wish to see ourselves versus how we are) or any need for maintaining an image for others (Gadanho & Lhuillier, 2007). After all, *social desirability* and *self-presentation* have been found to affect the answers people give e.g. in psychological research situations (Schwarz, 1999).

Another aspect related to trust is that users need to have some level of trust in the system in order to be ready to divulge personal information to it (Pu, Chen, & Hu, 2012). Bigger players, e.g. Amazon.com, in e-commerce profit from this, since users tend to trust them to keep their information confidential (Pu, Chen, & Hu, 2012).

Although typically implicit and explicit approaches to preference data gathering are easily seen as an either-or choice, in fact the range from implicit to explicit is a continuum and the degree of user involvement varies in different techniques and methods (Pommeranz et al., 2012). Moreover, explicit and implicit approaches can be combined to reduce effort without reducing quality (Xiao & Benbasat, 2007) and it has even been suggested that combining the results of explicitly and implicitly elicited preference data improves recommendation accuracy (Knijnenburg et al., 2012).

In practice, implicit approaches allow a great amount of data to be collected while motivating users to provide explicit data can be challenging, and the sheer amount of data allows implicit data, despite its noisiness, to be used to produce high-quality recommendations (Harper et al., 2005; Ling et al., 2005; Gadanho & Lhuillier, 2007; Konstan & Riedl, 2012). Consequently, while many academic studies are based on explicit recommendation data, e-commerce players tend to use large amounts of implicit data (Konstan & Riedl, 2012). In effect, neither approach (or the type of data they produce) is inherently better or more correct, as both come with their challenges and limitations in addition to their strengths.

Another reason for using data from both is that preferences collected implicitly and explicitly, in general, do not necessarily lead to the same conclusion: "the two methods are *not* equivalent and using one method over the other may result in a recommendation that does not match the preferences of the consumer" (Aggarwal & Vaidyanathan, 2003). As both approaches are subject to various biases, this is probably not all that surprising (Aggarwal & Vaidyanathan, 2003). Overall, inconsistencies between the implicit and explicit data in the study by Aggarwal and Vaidyanathan (2003) were fewer for aggregate data and the results were also more consistent for the more important item attributes. We do not yet know which method results in "truer" data (Aggarwal & Vaidyanathan, 2003)—and this may depend on such aspects as domain (Pommeranz, Wiggers, & Jonker, 2011), elicitation technique (e.g. rating vs. tweaking), and other contextual factors—but an assumption can be made that "true" preferences are somewhere in the middle (Aggarwal & Vaidyanathan, 2003).

Combining implicit and explicit data also allows the recommender system to move beyond the say-do problem; in a sense, implicit data is based on behavior and action, i.e. doing, while explicit data is based on saying. By combining the two, we can perhaps get a little closer to the true preferences that are palatable to the user—as discussed, sometimes the truth, the whole truth, and nothing but the truth, may not be in line with our self-image or the public image that we wish to project.

In addition to preference data and demographic data, today recommender systems are also increasingly gathering social information—followers and followed, tags, friend graphs, and so forth—and there is a clear tendency towards information from the Internet of Things, e.g. GPS location data, RFID, and real-time health signals (Bobadilla et al., 2013). However, additional research is still needed for us to be able to use all the available data effectively in forming recommendations (Konstan & Riedl, 2012).

## 7.2 IMPACT OF EFFORT

Implicit inputs do not involve user effort beyond using the service, while explicit inputs involve user effort (Gretzel & Fesenmaier, 2006; Gadanho & Lhuillier, 2007; Buder & Schwind, 2012). In general, cognitive effort is seen as something users try to avoid or reduce, and, consequently, asking a user to provide information for the sake of personalization is often argued to place an undesirable burden on the user (Gretzel & Fesenmaier, 2006; Pu, Chen, & Hu, 2012). In fact, Pu, Chen, and Hu (2012) give as guideline that designers should "[m]inimize preference elicitation in profile initiation" because humans prefer to save effort even at the cost of lower accuracy.

However, the equation is not that simple. For example, the *effort-accuracy model* conjectures that effort reduction is not an automatic goal; the goal is "to reach a satisfactory compromise between amount of effort and decision accuracy" (Gretzel & Fesenmaier, 2006). In fact, it has been found that people are willing to exert effort when they perceive the effort to lead to a good decision (Gretzel & Fesenmaier, 2006) or useful recommendations (Cremonesi, Garzotto, & Turrin, 2012b). Effort, in fact, appears to be used as a heuristic for quality especially when the quality of the object to be evaluated is difficult to ascertain (Gretzel & Fesenmaier, 2006). Also, exerting effort leads to more confidence in the decision (Gretzel & Fesenmaier, 2006).

In fact, already in 2002, a user-study by Sinha and Swearingen showed that users perceive systems that required more ratings as high on satisfaction and usefulness. Also, time to register and receive recommendations did not correlate with perceived usefulness (Sinha & Swearingen, 2002). Consequently, Sinha and Swearingen (2002) concluded that:

> "Ultimately what mattered to users was whether they got what they came for: useful recommendations. Users appeared to be willing to invest a little more time and effort if that outcome seemed likely."

This is, of course, very much in keeping with what the effort-accuracy model posits (Gretzel & Fesenmaier, 2006).

Also, when comparing different ways of giving preference input (ranking, ordering, and navigational), Pommeranz et al. (2012) found that the effort exerted in the elicitation phase does not affect how well an approach is liked. In fact, effort decreased liking only when the compared tasks were similar in terms of the process and type of input—rather obviously, exerting more effort on a comparable task is not liked—but when the process and the type of feedback were different, effort was not a predictor of liking or not liking.

Likewise, Gretzel and Fesenmaier (2006) found that increasing the number of questions asked had only a marginal impact on perceived enjoyment but increased perceived value. Greater effort seemed to lead to greater expectations, however, and if a suggested item, in this case a spring break travel destination, was not a good fit to the user's needs, evaluations were slightly less favorable. In effect, greater effort led to increased expectations concerning personalization and resulted in users being more aware of mismatches between the recommendations generated and their needs. Interestingly, the study also suggests that the relevance of the questions asked in the preference elicitation stage has very little impact on perceived value.

Moreover, Cremonesi, Garzotto, and Turrin (2012b) set out to study "the tradeoff that exists between *maximizing* the *user utility* and *minimizing* the rating effort", i.e. to see whether it was *user utility* or *user effort* that affected the perceived quality of user interaction more strongly. They used the profile length—how many ratings users must make before receiving recommendations—as the measure of effort. They found, as did numerous other studies that they quoted, that "[i]f a more demanding rating process is balanced by significantly better recommendations, the global satisfaction is not affected negatively by the increased effort." It appears that the two contrasting aspects, accuracy and effort, compensate each other (Cremonesi, Garzotto, & Turrin, 2012b).

Consequently, there is a trade-off between effort and accuracy, and therefore asking questions beyond the threshold of when additional questions no longer provide added value is discouraged. In effect, effort appears to function as an important cue for quality of the recommendation. (Gretzel & Fesenmaier, 2006)

Cremonesi, Garzotto, and Turrin (2012b) also found by testing with a collaborative filtering algorithm (PureSVD that is based on matrix-factorization and has been shown to be "one of the best" as far as accuracy is concerned) and a content-based algorithm (DirectContent, a simplified version of the LSA algorithm) that, between five and ten ratings, the perceived relevance of the recommendations rose with the number of ratings, but between ten and twenty ratings, the perceived relevance *decreased*. Therefore, it appears that this is where the added effort no longer provides enough increased accuracy to justify the spent effort—the positive force of increased relevance is overcome by the negative force of increased effort. Cremonesi, Garzotto, and Turrin (2012b) consequently distilled their result into a heuristic that "10 ratings are enough" for perceived relevance and that there is no need to collect extremely long profiles.

Similarly, Drenner, Sen, and Terveen (2008) had previously concluded that new users only need to enter five ratings for good enough accuracy based on replaying the history of a random sample of five thousand users (selected from users with more than 65 ratings and representing about 30% of such users) from MovieLens data. They felt that the user experience would suffer from having to make more ratings before getting recommendations, given that mean absolute error (accuracy measure) increased only by 0.073 when using five ratings instead of fifteen ratings. However, they did not give this number (five ratings) as a recommendation to other systems—they were trying to find a suitable number for ratings for an experiment—and they did not actually measure the user experience.

However, while "ten ratings is enough" is an important contribution, it must also be remembered that users do not only rate items in order to get more accurate recommendations (Herlocker et al., 2004; Harper et al., 2005). Users also rate items for *self-expression* (some power users of MovieLens who had rated over 1,000 movies stated that they rated items because it felt good), because it is fun, and to keep a personal list of movies they had seen (Herlocker et al., 2004; Harper et al., 2005). While ten may be enough, depending on the domain and context, more may be better.

In e-learning, especially if compulsoriness is used to elicit preference data, lowering effort may lead to increased dishonesty (i.e. rating items without reading them) and suspicions of dishonesty in the community-at-large, while increased effort can improve both honesty and user experience (Leino, 2012a; Leino, 2013; Leino & Heimonen, 2013). Consequently, the appropriate level of effort is domain-dependent and finding it continues to be a challenge for recommender system designers.

## 7.3 CONSTRUCTING PREFERENCES

Xiao and Benbasat (2007) feel that in eliciting user preferences, there tends to be an implicit assumption that users "recognize their own needs or at least have the ability to understand and answer the preference elicitation questions". This may not always be justified, as users may be dealing with items about which, or about the use of which, they do not have the required knowledge to specify their preferences appropriately (Xiao & Benbasat, 2007). Moreover, Xiao and Benbasat (2007) suggest that users, in fact, often do not answer the question asked but rather the question they *thought* was asked.

There is another implicit assumption in play that also affects the preferences that users input: It is assumed that users, in fact, have stable preferences that they can provide when asked and that the results of the elicitation process can be trusted as true and authentic user preferences (Gretzel & Fesenmaier, 2006; Xiao & Benbasat, 2007; Pommeranz, Wiggers,

& Jonker, 2011; Zhang, 2011; Kluver et al., 2012). However, there are various views concerning how preferences are formulated and what happens when they are measured (Kluver et al., 2012; Pommeranz et al., 2012).

At the one end of the spectrum is the idea of *articulated values* that holds that humans have well-reasoned, stable preferences for most things that are within the scope of their experience, and variance in elicited values[32] is seen as coming from "subtle differences in understanding of questions". At the other end of the spectrum, there is the idea of *basic values* that posits that humans only hold a store of "a small number of opinions or beliefs" and they use these to "derive momentary preferences when needed". In this train of thought, variance comes "from priming effects, mood, and other psychological phenomena affecting the process used to derive preference." (Kluver et al., 2012; see also Vermeir, Van Kenhove, & Hendrickx, 2002)

Overall, there appears to be a growing consensus that preferences tend to be ill-defined and constructed on the spot when needed in the decision-making context (Schwarz, 1999; Xiao & Benbasat, 2007; Pommeranz, Wiggers, & Jonker, 2011; Pommeranz et al., 2012). In constructing attitude judgments, we draw on the information that is most accessible at that particular point in time (Schwarz, 1999). Some information is *chronically accessible*, i.e. always accessible when we think of the topic, and some is *temporarily accessible*, i.e. comes to mind because of contextual influences (e.g. the rating scale used—see Section 7.5 for details—or the preceding or subsequent question) (Schwarz, 1999). For our mind, especially in low-involvement situations, the information that is not activated "might as well not exist"—a situation Kahneman (2012, p. 46, 85–88) describes with the abbreviation WYSIATI: What You See Is All There Is. Moreover, as soon as enough information has been retrieved to enable making the judgment with sufficient subjective certainty, the information search process is truncated and the judgment is made based on that subset of potentially relevant information, i.e. the most accessible information (Schwarz, 1999).

In addition, our reasoning is accompanied by *metacognitive experiences* that influence our judgments in addition to and even at the expense of declarative information—unlike assumed in most theories of human judgment, we do not form judgments purely based on accessible declarative information; instead, judgments emerge from the interplay of declarative and experiential information. In words of Schwarz (2004), "there is more to thinking than thought content." The two most notable metacognitive experiences are 1) *accessibility experiences*, or "the ease or

[32] For example, when users are asked to re-rate items, the new ratings can differ significantly from the previous ones (Cosley et al., 2003; Amatriain et al., 2009).

difficulty of recall and thought generation," and 2) *fluency experiences*, or "the fluency with which new information can be processed." These subjective experiences are informative in and of themselves and are used in making decisions and judgments. (Schwarz, 2004)

By and large, when recalling is experienced as easy, we tend to judge consistently with the recalled content, and when recalling is experienced as difficult, we tend to go against the implications of the recalled content (unless we have a justifying reason for the experienced difficulty, e.g. lack of expertise or presence of distractions). For example, when participants were first asked to recall six examples of either assertive or unassertive behavior and then asked to rate their own assertiveness, those who had recalled six examples of assertive behavior unsurprisingly rated themselves as more assertive. However, when the experiment was replicated so that participants were asked to recall 12 examples, the pattern was reversed—participants who successfully recalled 12 examples of unassertive behavior rated themselves as *more* assertive than those who had recalled 12 examples of assertive behavior. It appears that while it is relatively easy to recall six instances of (un)assertive behavior, recalling 12 instances is experienced as difficult—and if it is difficult to recall 12 instances of (un)assertive behavior, then the participant concluded that they cannot be that (un)assertive[33]. Importantly, the effect of accessibility experience is mediated by motivation: Low processing motivation tends to lead to using subjective accessibility experiences in making decisions and judgments (a heuristic processing strategy), while high processing motivation tends to lead to emphasizing accessible content (a systematic processing strategy). Also, the mood affects the equation, as happy mood promotes reliance on heuristic strategies (e.g. using accessibility experiences), while sad mood promotes reliance on systematic strategies (e.g. using accessible content). (Schwarz, 2004)

The experience of the fluency of processing functions as an experiential basis of truth judgments—the more fluent the processing is experienced to be, the more truthful the information appears to us. Moreover, fluent processing also leads to the feeling of having known "the truth" all along. It appears that this effect is based on familiarity—if we feel that we have heard it before, we presume that there is something to it. In effect, this illusion of truth based on the frequency of exposure has been demonstrated with wartime rumors, foreign language words, and advertising materials, and it has persisted even when participants have

[33] In fact, the experiencing of ease or difficulty can also be caused with different means with the same result. When participants were asked to recall six instances of assertive behavior but one group was induced to contract the corrugator muscle, i.e. to furrow their brow (associated with effort), and the other group to contract the zygomaticus muscle, i.e. to produce a light smile (associated with a feeling of ease), the furrowing group rated themselves as *less* assertive than the smiling group. This underlines further that the subjective experience of ease or difficulty affects our judgments.

been told at the time of the exposure that the information is false. The experience of processing fluency influences judgments of liking, preference, and even beauty, suggesting that processing fluency is experienced as inherently positive. Repeated exposure even to a neutral stimulus results in gradual increase in liking (so-called *mere exposure effect*) and even single exposure can increase liking. In effect, perceptions of beauty and truth are both enhanced by processing fluency[34]. Significantly, processing fluency also leads to spontaneous affective reaction, as high fluency results in stronger activity in the zygomaticus region, i.e. "smiling muscle," that is associated with positive affect. In effect, processing fluency influences human judgments through the fluency experience itself and by the spontaneous affective reactions that it elicits. Moreover, the relationship between familiarity and affective response is bidirectional, as is has been demonstrated that stimuli that elicit a positive affective response are also judged to be more familiar. Significantly for eWOM, it has been demonstrated that the fluency with which the product information can be processed affects the decision process—a font that is hard to read can lead to the product being seen as less attractive. Consequently, facilitating processing fluency also in presentation is crucial. (Schwarz, 2004)

This leaves the constructed preferences subject to the influence of the context and biases arising from it that can be substantial (Schwarz, 1999; Gretzel & Fesenmaier, 2006; Xiao & Benbasat, 2007; Zhang, 2011; Adomavicius et al., 2013). Customer preferences have been found to be susceptible to various, seemingly irrelevant factors, e.g. the set of alternatives and the way questions are asked (Gretzel & Fesenmaier, 2006; Pommeranz et al., 2012), leading to preference reversal when preferences are elicited with different tasks, e.g. choice vs. rating task (Xiao & Benbasat, 2007). In effect, it appears that the preferences elicited by recommender systems are not just the function of the match between user preferences and the suggested item but also a reaction that the features of the system and the method or process of elicitation influence (Gretzel & Fesenmaier, 2006; Pommeranz et al., 2012). Consequently, recommender systems can already have a persuasive influence during the preference elicitation process (Gretzel & Fesenmaier, 2006).

Moreover, humans are not entirely rational—we are boundedly rational, as our rationality is limited by our information processing capacity and other factors, such as availability of information (Simon, 1955; Xiao & Benbasat, 2007). Moreover, decision-making is also affected by emotional factors (Damasio, 2006, xiv-xvii); e.g. affect influences what is recalled first (Pommeranz et al., 2012) and mood causes variance in preferences (Kluver et al., 2012).

[34] As Schwarz (2004) points out, this underlines the percipiency of Keats's assertion that "beauty is truth, truth beauty."

In effect, preferences are not something readily available in the memory (Schwarz, 1999; Gretzel & Fesenmaier, 2006) and, consequently, the preference elicitation interfaces and processes need to be studied further for us to be able to elicit correct and accurate preference data from users (Pommeranz et al., 2011).

## 7.4 PREFERENCE ELICITATION PROCESS AS A FORM OF DISCOURSE

As discussed in Section 4.2 (discussion on Persuasion), users perceive recommender systems as social actors and ascribe personalities to them. However, preference elicitation process is typically seen as a way of learning user preferences without considering the social dimensions of the process that "is colored by attitudes, judgments, stereotypes, and affective reactions" (Buder & Schwind, 2012; see also Gretzel & Fesenmaier, 2006). Based on their study, Gretzel and Fesenmaier (2006) conceptualize user-recommender relationship as quasi-social, meaning that user "interaction with recommender systems have to be understood as a form of discourse" (see also McNee, Riedl, & Konstan, 2006b). Asking questions is, in effect, inherently a social process (Gretzel & Fesenmaier, 2006).

Given the constructive nature of human information processing, the cues that recommenders provide in the course of the interactive preference elicitation process affect preference construction (Gretzel & Fesenmaier, 2006; Zhang, 2011; Adomavicius et al., 2013). In other words, recommenders persuade and affect the preferences they aim to elicit by the very questions they pose (Gretzel & Fesenmaier, 2006; Zhang, 2011). For example, anchoring and adjustment heuristic[35] tends to lead to biases in judgment (Zhang, 2011), and, given that anchoring appears to operate at the time that the user is constructing the response (Zhang, 2011), it is no wonder that e.g. Zhang (2011) and Cosley et al. (2003) have shown that interface cues, e.g. ratings, affect the rating that the user makes.

Given the fact that preference elicitation process typically comes at the beginning of the interaction with recommenders, meaning that it colors the rest of the interaction, it is no wonder that it tends to establish the personality of the recommender for the user and that it constitutes "an intrinsic factor that ultimately determines a user's evaluation of the recommendation provided" (Gretzel & Fesenmaier, 2006; see also McNee, Riedl, & Konstan, 2006b).

[35] *Anchoring and adjustment heuristic* refers to our tendency to rely on the first piece of information that we get—the anchor—when making subsequent judgments and decisions. In effect, we make judgments by adjusting from the anchor and, therefore, the subsequent judgments tend to cluster around the anchor; e.g. seeing a predicted rating for an item tends to affect our rating of the item (Cosley et al., 2003; Zhang, 2011). For further information on the heuristic, see e.g. https://en.wikipedia.org/wiki/Anchoring_and_adjustment.

## 7.5 Rating Items

As Web 2.0 has turned users into content providers, almost all social applications nowadays provide users opportunities to rate items either for personalization or social purposes (Kuflik et al., 2012). Item-rating—i.e. *users* assigning *ratings* to *items*—is a common strategy for recommender systems to model users (Cosley et al., 2003). Item-rating is important in the context of this dissertation because, besides being perhaps the most common explicit approach in use, most of the publications that constitute this compound dissertation deal with item-rating or item-rating-related matters.

Schafer et al. (2007) suggest that it will still take decades before information systems can be expected to automatically recognize such subtle information as aspects of aesthetic taste that are important to humans, and, until then, humans need to be included in the loop to analyze information and to condense their opinions into a form of data that is easy to process. Ratings are crucial for recommender systems to be able to provide accurate, useful recommendations (Redpath et al., 2010; Gena et al., 2011; Pommeranz et al., 2012). At the same time, as discussed above, there are numerous factors that affect the reliability of the ratings that users provide, and being able to elicit correct ratings should not be taken for granted but be designed and evaluated carefully (Gretzel & Fesenmaier, 2006; Pommeranz, Wiggers, & Jonker, 2011; Sparling & Sen, 2011; Zhang, 2011; Adomavicius et al., 2013)—in actuality, "user ratings are imperfect and noisy" (Kluver et al., 2012).

#### Item and Domain Characteristics

Different items are rated in different domains, and, consequently, different domains may suggest different rating scales, as the effectiveness of the rating scales appears domain-dependent (Cosley et al., 2003; Sparling & Sen, 2011). Sparing and Sen (2011) discussed three domain-related characteristics:

1) *Experienced vs. remembered*: Some items are experienced offline and attitudes towards them are recalled when they are rated, while some are experienced on the same page basically at the same time that they are evaluated. For example, a book read is recalled when rated but a book review is read on the same page that it is rated.
2) *Rating distribution*: The overall distribution of the rating values is affected by the domain; e.g. YouTube switched from a five-star scale to binary one (Thumbs-up/down) because the middle values, 2-4 stars, constituted only about 5% of the ratings.
3) *Agreement*: Some domains are more conducive to agreement than others; in them, users are more or less in agreement about the items. For example, users agree more typically about customer reviews on Amazon.com than about movies on MovieLens.

As discussed in Section 6.3 (discussion on Item Type), in e-commerce items—typically goods and services—can be divided into three types: Search, experience, and credence goods and services. *Search goods/services* are the goods and services for the quality of which users can easily obtain information prior to purchase, *experience goods/services* are goods and services that need to be sampled or consumed to be able to determine their quality, and *credence goods/services* are goods and services that cannot be evaluated for quality even after consuming them; e.g. a haircut (Yang & Mai, 2010; Mudambi & Schuff, 2010; Racherla & Friske, 2012). While some goods or services may represent quite pure examples of one of the types, in many cases goods and services have various attributes that fall into different types, and so goods and services "can be placed on the continuum of search, experience, and credence attributes" (Yang & Mai, 2010; see also Racherla & Friske, 2012).

Consequently, these different types of item attributes can give arise to different types of evaluations, and rating an item that has a lot of credence attributes can be challenging. However, while the differences in the type of items have been studied as far as eWOM is concerned, their effect on rating agreement or rating distribution appears not to have received much research interest.

However, we do know that the item to be rated influences the rating (Gena et al., 2011; Kuflik et al., 2012) and therefore recommender algorithms should not be evaluated with a data set coming from a very different domain; the tasks that the algorithm is designed to support should be similar to the tasks in the domain from which the data comes (Herlocker et al., 2004).

#### Rating Scales

Choosing the rating scale is a major decision in recommender system design (Cosley et al., 2003). Rating scales allow users to express their preferences (Gena et al., 2011).

Schafer et al. (2007) defined rating as consisting of an association of two things, *user* and *item*, typically by means of a *value*. The value is what the rating scale allows the user to assign to the item. The value can be a *scalar rating* on a numeric, e.g. 1-5 stars, or ordinal scale, e.g. strongly agree, agree, neutral, disagree, strongly disagree. A *binary rating* allows choosing between two positions, e.g. yes/no, agree/disagree, or good/bad. Finally, a *unary rating* can only indicate one thing, e.g. that a user has viewed or purchased the item or has rated it (typically) positively, e.g. Facebook.com's "like"[36]—absence of the unary rating typically only tells

[36] Theoretically, of course, there could be a unary scale that only allowed taking a negative position in relation to the item but we are not aware of any such system.

us that we have no information as far as the user and the item are concerned; e.g. the user may have viewed or purchased the item elsewhere. (Schafer et al., 2007)

Gena et al. (2011) define rating scales as *complex widgets* that are characterized by i) granularity, ii) numbering, iii) visual metaphor, and iv) neutral position:

- *Granularity* refers to the number of positions on the scale and can be coarse, e.g. binary scale, or fine, e.g. five stars at half-a-star steps (i.e. 10-point scale).
- *Numbering* refers to the numbers, if such exist, that can be associated to each position on the scale; e.g. a scale with three positions might be numbered for example 1, 2, 3 or -1, 0, 1.
- *Visual metaphor* refers to the visualization form that the scale implementation on the interface takes, e.g. a slider, thumbs-up and down icons, stars, etc. Visualizations convey metaphors and have emotional connotations, e.g. a slide implementation is based on a technological metaphor and appears typically as cold or detached. (Cena, Vernero, & Gena, 2010; Gena et al., 2011)
- *Neutral position* refers to the presence or absence of an intermediate neutral point on the scale that allows users to indicate that they have no opinion.

All of the above features of rating scales contribute to what Gena et al. (2011) call the *personality* of the rating scales, i.e. "the way ratings scales are perceived by users and affect their behavior."

An ideal rating scale allows users to easily express their preferences in a meaningful way, i.e. scales should allow users to make *meaningful* distinctions (Cosley et al., 2003). In other words, when asking users to map opinions to a number between 1 and N, N should be high enough that users can distinguish between as many different levels of liking as they would like to but not too high so as to leave users unable to make judgments between the levels (Cosley et al., 2003; Herlocker et al., 2004). In fact, psychological literature indicates that simply expanding the number of choices does not automatically increase scale sensitivity, underlining that how users respond to different rating scales is rather a psychological issue than a mathematical question (Kuflik et al., 2012). In addition, the scale naturally needs to allow the system to use the preference data efficiently, e.g. the system should be able to use it to make accurate predictions (Cosley et al., 2003).

Mapping opinions to ratings is challenging, as opinions can be complex and the granularity of the scale may be different from the granularity of the user preference, e.g. the scale asks the user to rate an item on a scale from 1 to 5, while the user might only care if the item is good or bad (Cosley et al., 2003; Herlocker et al., 2004).

In any case, it may be difficult to people to express their preferences numerically in the first place in many contexts. In effect, there may be a mismatch between the quantitative preference representation in the recommender system model and the user's mental preference model; e.g. how much more would you like a 30-day holiday than a 28-day holiday—1.2-times more or perhaps 2-times more? Domain expertise also affects the user's ability to make distinctions and transfer them to numerical attribute weights. (Pommeranz et al., 2012)

In practice, the rating scales in use vary widely in granularity; e.g. Facebook's unary "like" button, Amazon's five-star scale, TiVo.com's -3 to +3 scale, MovieCritic's 14-point scale, and Launch.com's 0-100 scale for rating songs (Cosley et al., 2003; Redpath et al., 2010).

It appears that, other things being equal, finer-grained scales allow systems to generate better recommendations (Cosley et al., 2003; Sparling & Sen, 2011; Kluver et al., 2012). Also, users appear to like them better, preferring the middle of the granularity spectrum of scales the best (Gena et al., 2011; Sparling & Sen, 2011). In fact, the five-star scale is a clear user favorite (Cosley et al., 2003; Gena et al., 2011; Sparling & Sen, 2011). Given that being able to rate movies at half-stars was the most requested feature-change in MovieLens (Cosley et al., 2003), it appears that scales from 5 to 10 points are the ones that users prefer (Gena et al., 2011; Sparling & Sen, 2011). Sparling and Sen (2011) conjecture that finer-grained scales are preferred, especially in the domains where item-evaluations are more subjective. Also, more granular scales might also serve to reduce noise in rating data (Kluver et al., 2012).

However, as the scale granularity increases, users end up rating fewer items (Sparling & Sen, 2011; Leino, 2013). This seems connected to finer-grained scales resulting in longer rating times (Sparling & Sen, 2011). In effect, users are able to faster rate items that they really like or dislike, and it is the middle areas of the finer-grained scales where ratings take more time (Sparling & Sen, 2011).

#### Rating Scales Affect Ratings

Rating scales appear to influence the ratings that users give—items are rated differently depending on what kind a scale is used (Cosley et al., 2003; Cena, Vernero, & Gena, 2010; Gena et al., 2011; Kuflik et al., 2012). Various factors appear to be behind the biases that rating scales introduce into the ratings (Gena et al., 2011). For example, changing the range of a scale can affect the distribution of ratings (Cosley et al., 2003).

In psychological research, it has been found that one way that rating scales affect the ratings given is that people draw on the apparently formal features of the scale, such as its numeric values, to make sense of the questions they are being asked. In other words, rating scales unintentionally provide informational value to the raters. For example, Schwarz (1999) tells how changing the numeric scale dramatically changed the answers the respondents gave to the question "How successful would you say you have been in life?" Using the first scale, ranging from 0 ("not at all successful") to 10 ("extremely successful"), only 13% of the representative sample reported high success. In contrast, using the second scale, ranging from -5 ("not at all successful") to 5 ("extremely successful"), 34% reported high success. The difference was due to how the term "not at all successful" was interpreted. On the first scale (0 to 10), zero was seen as reflecting absence of great achievements but on the second scale (-5 to 5), -5 was understood to reflect explicit failures. (Schwarz, 1999)

The provided set of response alternatives appears to constitute a source of information also in other ways. Respondents tend to assume that the scale is constructed meaningfully, i.e. based on the knowledge of or expectations about "the real world." Consequently, especially in case of questions that concern activity frequency, the middle part of the scale tends to be seen as reflecting average, usual, or normal behavior while the extremes are seen as corresponding to more extreme, or less frequent or usual, behavior. Schwarz (1999) describes how in a study done in 1985, only 16.2% of the respondents admitted watching TV for more than two and a half hours a day when given a scale with low-frequency response alternatives, while 35% reported spending more than two and a half hours a day watching TV when given a scale with high-frequency alternatives. (Schwarz, 1999)

The TV watching study provides also an example of how previous questions can color the answers to subsequent questions. The respondents who reported watching two and a half hours of TV a day on the low-frequency scale (where it created an impression that they were watching more TV than most) were less satisfied with what they did on their leisure time than the respondents who reported watching two and a half hours of TV a day on the high-frequency scale (and therefore had an impression of watching less TV than most). In effect, the scale was used as a frame of reference when comparative judgment was subsequently made—the scale provided the respondents comparison information about themselves in relation to others that was used in the subsequent judgment. Therefore, the contextual influence of the rating scale used may result in both comparative and non-comparative evaluative judgments that do not reflect the assessments that the respondents would make in daily life. (Schwarz, 1999)

The order of questions can also influence which considerations are used in making the judgment. When respondents were first asked how satisfied they were on their life as a whole and then how satisfied they were with their marriage, the correlation between the judgments was $r$ = .32. However, when the order was reversed, the correlation rose to $r$ = .67. In effect, the marital satisfaction question brought marriage-related information to respondents' mind and thereby influenced their judgment of their life satisfaction more strongly than when this information was not first brought to their mind. This is known as *assimilation effect* where the positivity or negativity of the information brought to mind affects the subsequent judgment in the same way (positivity results in a more positive judgment and vice versa). The opposite effect is known as *contrast effect* where the information brought to mind becomes a standard against which something is evaluated. A positive standard of comparison results in a more negative judgment and vice versa. In effect, the same information can shift evaluations to opposite directions, depending on how it is used—whether it becomes a mental representation of the target (assimilation) or a standard against which the target is evaluated (contrast). (Schwarz, 1999)

Moreover, the idea of the center part of the scale representing normal or usual can also influence the answers due to *social desirability* and *self-presentation* (Schwarz, 1999). What if the scale appears to show us as somehow deviating from normalcy or acceptable? Do we want to report this and do we want to see ourselves in that way? Survey researchers try to reduce the impact of this by emphasizing anonymity and confidentiality (Schwarz, 1999). However, in case or recommender systems, the responses have to be connectable to individual respondents for the system to be able to use them for generating predictions.

In effect, rating scales are not neutral measuring instruments; instead, they have informational value and thereby can and do bias and influence the responses by functioning as a frame of reference for evaluative judgments. In effect, the same question with different rating scale may result in a different assessment. (Schwarz, 1999)

The influence of various rating scales has also been studied within the recommender system research community. Overall, it appears that the rating scales that convey a human-related metaphor, e.g. thumbs-up/down or smiley-face-based scales, tending to push ratings up while the presence of numbers tends to result in lower ratings (Kuflik et al., 2012). Also, more coarse-grained, i.e. lower-granularity, scales tend to promote higher ratings; it seems that at least some types of items are given the benefit of the doubt when users are forced to rate them on coarse scales (Cosley et al., 2003; Kuflik et al., 2012). For example, Cosley et al. (2003) noticed that in rating movies, the users who had given a movie 1-2 stars on a five-star scale, gave it thumbs-down on a thumbs-up/down scale, while the users who had given it 3-5 stars rated it with thumbs-up.

Also, scales with no zero, or neutral, position, can push ratings upwards (Cosley et al., 2003; Kuflik et al., 2012). In face-to-face interview situations, the reason for higher ratings has been suggested to be social desirability bias, i.e. wanting to please the interviewer or not to be seen as giving a socially unacceptable answer (Garland, 1991). The same may apply to online situations, too (Cosley et al., 2003; Gena et al., 2011; Kuflik et al., 2012). In effect, not to provide a neutral middle point is to force choice on users (Garland, 1991; Gena et al., 2011).

In addition, when the lower end of the scale is labeled with negative numbers (e.g. -2, -1, 0, +1, +2) rather than with positive numbers (e.g. 1, 2, 3, 4, 5), it is seen as more negative, and, consequently, ratings tend upwards (Gena et al., 2011; Kuflik et al., 2012).

The five-star scale may be an exception to the above-mentioned rules-of-thumb, as it tends to promote higher ratings (Kuflik et al., 2012). Consequently, Kuflik et al. (2012) hypothesize that the *popularity* of a scale also affects the ratings it yields, and should therefore be seen as contributing to the personality of a scale.

The choice of the rating scale can significantly affect the level of noise in the preference data, with more granular scales producing less noisy data, although the effect is limited by the inherent discriminability of the rated item (Kluver et al., 2012).

Many other aspects of the scale implementation in the interface can also affect the ratings that the scale produces, including the labels associated with choices, question interpretation, ratings scale balance (whether the negative and positive sides are in balance), and order of choices (there is some evidence of a bias towards the left side of the scale) (Cosley et al., 2003; Gena et al., 2011).

#### Rating Scale Personality

A group of researchers, e.g. Cena, Vernero, and Gena (2010), Gena et al. (2011), and Kuflik et al. (2012), has proposed that rating scales have their own personalities that influence users' rating behavior. They argue that the features of rating scales, granularity, numbering, visual metaphor, and neutral position, all "contribute to define … the personality of rating scales, i.e., the way rating scales are perceived by users and affect their behaviour" (Gena et al., 2011). In effect, Gena et al. (2011) argue that rating scales are not a neutral part of the rating equation but a factor that affects the ratings that users give, i.e. exert influence on users who use them to express their preferences. Consequently, user ratings are determined at least by three factors: The item being rated, the user rating the item, and the rating scales used (Gena et al., 2011).

We have already considered the effect of granularity to the rating behavior. Another point emphasized in this line of inquiry is that "each rating scale conveys a different metaphor, which influences the emotional connotation"; e.g. thumbs metaphor is related to human behavior and is friendly and young but also impolite and overly simple; stars rely on cultural conventions and are classical and familiar; and slides have a technological feel and are precise, cold, detached, and boring (Cena, Vernero, & Gena, 2010).

Gena et al. (2011) suggest that rating scale personality can be defined at two levels, aggregate and individual. At *aggregate* level, the rating scale personality affects the overall rating behavior, reflecting general propensities in how rating scales are perceived and used. This is the level that has been studied in the publications mentioned here. At *individual* level, each user may be influenced in an idiosyncratic manner by the rating scale personality.

In addition to the levels suggested by Gena et al. (2011), we propose that there can also be a third level, *sub-community* level, where the rating scale influences some sub-group of the community; e.g. given that thumbs appear youthful but impolite, it may be that younger users do not hesitate to give thumbs-down while older users may perceive it as impolite and consequently use thumbs-down only when the preference is especially strong.

Overall, we can take it as given that the rating scale is one of the determining factors in rating behavior, in addition to the item being rated and the user rating the item (Gena et al., 2011). However, we may wish to question if this translates into rating scales having distinct personalities. The evidence provided is not always objection-proof, e.g. while Kuflik et al. (2012) state that "the collected data confirm that that rating scales have a 'personality' which exerts an effect on user ratings", they continue that "[w]hile visitors in general favored the presentations they viewed, the average ratings differed (*in some cases the differences were statistically significant*) between the different rating scales" (emphasis added). Personality implies something relatively permanent and consistent. Are rating scale personalities consistent over domains or do domains moderate their effect? There is a number of questions concerning whether or not we should talk about rating scales having personalities, but the suggestion is nevertheless both intriguing and worth further research.

#### Converting Ratings between Scales

There have been suggestions that users should be allowed to rate items on the scale of their choosing, as different users appear to have different preferences as to rating scales (e.g. Cosley et al., 2003; Gena et al., 2011). Moreover, there are also other situations where moving from one scale to another is necessary, e.g. a researcher may need to compare data that has been collected using different scales (Kuflik et al., 2012). However, this begs the question of how to convert the ratings from one scale to another. Given the discussion above of how rating scales cannot be seen as neutral tools, the answer is not simple.

Cosley et al. (2003), based on re-ratings on various scales, concluded that "user ratings correlate very well between scales," and so recommendations can be computed "using normalized scores." In fact, normalization is not an uncommon procedure in calculating recommendations in any case (Redpath et al., 2010; Konstan & Riedl, 2012). Hill et al. (1995) previously had similar results when testing how consistently users re-rated items after six weeks of making the original ratings. Cosley et al. (2003) tested to see if a similar result could be had when months or years had passed since the original rating. While the correlation was lower than in Hill et al. (1995), the correlation was, as mentioned, still robust (Cosley et al., 2003).

In contrast, Cena, Vernero, and Gena (2010) found that "40% of ratings depart considerably from mathematical proportion, showing that mathematical proportion is not enough to make a mapping which is able to capture the actual meaning of user ratings." Kuflik et al. (2012) similarly conclude that "given the different distributions no linear transformation can exist."

Gena et al. (2011) and Kuflik et al. (2012) consider this to mean that this challenges the conclusions of Cosley et al. (2003). However, this seems to be a discussion of apples and oranges, as Cosley et al. (2003) also pointed out that different rating scales did affect the ratings (thereby showing false the claim in Kuflik et al. (2012) that Cosley et al. (2003) "implicitly" assumed "that rating scales are neutral tools which do not have any influence on the user ratings themselves") and did not claim that converting between scales can be done using a "pure mathematical solution", as claimed by Kuflik et al. (2012), but that ratings correlated well enough that good *recommendations* can be computed using *normalized scores*. The works of Cena, Vernero, and Gena (2010), Gena et al. (2011) and Kuflik et al. (2012) show that rating scales affect ratings, as does the study by Cosley et al. (2003), but they do not conclusively show that the fact that "40% of the rating departed considerably from mathematical proportion" (Cena, Vernero, & Gena, 2010) actually means that reliable recommendations could not be calculated from normalized scores.

In a sense, to show that would mean that using a different scale could "make a user rate a 'bad' movie 'good'", to use the words of Cosley et al. (2003). Cosley et al. (2003) showed that such switch is possible in some situations if the interface deliberately misleads the user. However, it has not been shown this far that this can be accomplished by switching to another rating scale. Changes in distributions observed by e.g. Kuflik et al. (2012) are still a far cry from this.

Converting between scales is undeniably complicated by the fact that rating scales affect rating behavior. Gena et al. (2011) conclude that "the coefficients for capturing the actual personality of rating scales should be learnt by users' behavior with a specific system, and cannot be calculated at [sic] priori."

#### Some Other Factors That Affect Ratings

There are various other factors that also affect the ratings that users provide. *Users* are naturally part of the equation that determines the final rating. Users can attribute different meanings to the same position on the same rating scale (Cena, Vernero, & Gena, 2010), e.g. three-stars on a five-star scale might mean "not very good but acceptable" to one user and "not good, in fact, unacceptable" to another. Also, the mood of the user at the time of the ratings influences the rating as does the relevance of the item to the current user task and the importance of the task (Cosley et al., 2003).

In addition, users have idiosyncratic behaviors, e.g. some may only provide positive ratings while others may use the full scale (Shardanand & Maes, 1995; Cena, Vernero, & Gena, 2010; Gena et al., 2011). In effect, there appear to be "behavioral tendencies to disproportionately select a subset of the available response options," i.e. there are systematic individual differences as far as response location and spread are concerned that cause a systematic bias across items in a questionnaire (Weijters, Geuens, & Schillewaert, 2010). In behavioral sciences, two response styles have been subject to large research interest, namely *acquiescence response style* (disproportionate use of positive options) and *extreme response style* (disproportionate use of extreme options) (Weijters, Geuens, & Schillewaert, 2010). Significantly, these response styles are independent of item content, meaning that they are different from social desirability effects (Weijters, Geuens, & Schillewaert, 2010).

The mechanisms that are available to users to rate items can also cause biases in the data set (Herlocker et al., 2004). For example, in an experiment on different methods to select items for first-time users to rate to initialize their profile, Rashid et al. (2002) demonstrated that each method they tested led to different biases in initial ratings.

Also, the recommender system *interface* affects the ratings in various ways, as "people are often influenced by elements in the environment in which preferences are constructed" (Adomavicius et al., 2013). One way is that the predictions shown in the interface influence the ratings users give to items; e.g. in re-rating movies, users tended to rate towards the prediction, whether it was accurate or not, even to the point of rating a "bad" movie "good" under some conditions (Cosley et al., 2003). The effect of the shown prediction may actually be strong enough to change opinions, although how long the effect lasts is uncertain (Cosley et al., 2003). As suggested by psychological literature on conformity, it appears that, while helping people make choices, recommender systems also affect their opinions (Cosley et al., 2003). This may, in fact, be related to the fact that users appear to view the recommendations "as a suggestion to a 'correct' answer"—users appear to "have some inherent trust in recommender systems" (Adomavicius et al., 2013).

Zhang (2011) found a similar *anchoring*, or *feel-forward, effect*, when studying anchoring effects during the construction of preferences at the point of consumption; in contrast, Cosley et al. (2003) had studied users re-rating items that they had previously consumed, and, consequently, there was a potential of recall-related effects. Like Cosley et al. (2003), Zhang (2011) also found "strong evidence that biased output from recommender systems can significantly influence the preference ratings of consumers" and that "the effect of perturbations on rating drift is continuous, not discrete"; i.e. not only extreme manipulations prompt the effect. Zhang (2011) argues that the anchoring effect operates at the time that users formulate their responses, as the effect was the same whether the system recommendation was seen before or after viewing a TV show, meaning that the effect is not attributable to priming (i.e. only before condition would show the effect). The results by Zhang (2011) also underline that the anchoring effect is not only in play in high uncertainty situations but also impacts rating when there is no need for recall and no uncertainty (Adomavicius et al., 2013)—after all, ratings in her experiment were given immediately after seeing the TV show.

Overall, the anchoring effect allows the effects of a successful shilling attack (Lam & Riedl, 2004) to multiply in the ratings—the impact of the attack is not limited to distortions caused by the dishonest rating(s) alone but is also enhanced by the effect the dishonest rating(s) have on the subsequent ratings of other users (Zhang, 2011; Adomavicius et al., 2013).

#### Noise in Ratings and the Magic Barrier

As a result of all the problems related to collecting user preference data, it is no surprise that the data that comes out of the process can be noisy, corrupted, or simply downright wrong (Kluver et al., 2012; Konstan & Riedl, 2012; Said et al., 2012). While most academic studies are based on explicitly collected preference data, in commercial practice it is much more typical to use implicitly collected preference data that is considered noisier than the explicitly collected data to begin with (Konstan & Riedl, 2012). Inconsistencies in the preference data, naturally enough, limit the recommender system prediction accuracy (Kluver et al., 2012; Said et al., 2012).

Noise in the preference data can be divided into *natural noise*, or noise resulting from the way the preference data is collected by the system, and *malicious noise*, or biased noise purposely introduced into the system to affect predictions (O'Mahony, Hurley, & Silvestre, 2006). Here we are concerned with natural noise and its effects. Natural noise comes from many sources, as discussed above.

The presence of natural noise in ratings, of there being "a random component" in addition to the "underlying true feeling of the user", led Hill et al. (1995) to speculations of there being an upper limit to prediction performance of recommender systems, according to "standard statistical theory … the square root of the observed test-retest reliability correlation." As Hill et al. (1995) observed a rating-rerating correlation to be 0.83 in the ratings that were six weeks apart, the theoretical preference prediction would therefore have a correlation of 0.91. The correlations observed by Cosley et al. (2003) after a much longer period of time between rating and re-rating were slightly lower, 0.70, meaning that the theoretical accuracy of the prediction would also have a lower correlation of 0.83 than the one in Hill et al. (1995) if we use the calculation approach by Hill et al. (1995).

The term *magic barrier* to describe the limit of accuracy that natural variability may prevent us from getting past was introduced by Herlocker et al. in 2004. Amatriain et al. (2009) define the magic barrier as "the asymptotic limit on accuracy improvements for all recommendation techniques, regardless of how fine-tuned they are."

There have been various approaches to reduce natural noise in ratings. Kluver et al. (2012) proved with synthetic datasets that ratings made on a higher-precision, i.e. finer-grained, scale improve predictive accuracy, but whether the same holds for natural datasets with different scales is still an open question. Amatriain et al. (2009) studied denoising datasets by asking users to rate some items again. While complete denoising was accomplished when each item was rated three times, this is naturally impractical in real use conditions. Consequently, they proposed two partial re-rating strategies, data- and user-dependent denoising. *Data-dependent denoising* strategy focused on re-rating only the extreme ratings (1-star and 5-star ratings, as the five-star rating scale was used) and required re-rating 20% of the ratings in their database, yielding over 5% improvement. *User-dependent denoising* strategy also achieved 5% improvement when less than 10% of the items in their dataset were re-rated by 30% of the users. Amatriain et al. (2009) also noted that at least in some circumstances, re-rating strategy yielded much higher accuracy gain than having users rate new and unrelated items.

The noisiness of the preference datasets also affects using the datasets to test different algorithms offline. User-centric approaches avoid this problem but come at the cost of time and having to have a set of users available for testing. The drawbacks mean that much of testing is still done using data-centric evaluation methods. (Said et al., 2012)

#### Summary

Given the importance of rating data to recommender systems and all the challenges related to the seemingly simple task of users rating items, it is surprising that there is little guidance available for designers in research literature for choosing which rating scale to employ in various settings (Sparling & Sen, 2011). In effect, in employing rating approaches in recommender systems, careful evaluation before employment is essential (Sparling & Sen, 2012). In any case, the idea that user ratings actually directly represent user preferences is rather questionable, as rating data has repeatedly been shown to be far from perfect, and so improving the data that recommender systems use to generate recommendations remains an active topic in research effort.

## 7.6 MOTIVATING CONTRIBUTING

Eliciting sufficient contributions from their users, i.e. the user community, is a life-and-death challenge to online communities, many of which wither to death due to under-contribution (Harper et al., 2005; Ling et al., 2005; Rashid et al., 2006). Also, the functioning and precision of community-based systems, such as recommender systems, is strongly dependent on the contributions of the users (Herlocker et al., 2004; Harper et al., 2005; Farzan & Brusilovsky, 2006). For example, in the MovieLens system, more than 20% of the movies do not have enough ratings to allow the recommenders to accurately predict whether users are going to like them or not (Ling et al., 2005; Rashid et al., 2006). In addition, even in the communities that do manage to elicit enough data to stay in the game, most of the contributions tend to come from a small fraction of the user community (Harper et al., 2005; Ling et al., 2005; Rashid et al., 2006). While not everybody needs to contribute to make the system successful, if a large part of the community does not contribute, recommender systems will face difficulties in trying to serve the community (Ling et al., 2005).

Although explicit feedback is typically seen as providing the most dependable information source for personalization , the fact that “users do not like to rate” has led to attempts to use increasingly implicit feedback instead (Farzan & Brusilovsky, 2006)—a common practice in e-commerce from the start (Konstan & Riedl, 2012). However, this has not provided a final solution to the problem (Farzan & Brusilovsky, 2006).

Consequently, motivating users to contribute is a vibrant branch of research in the field of recommender systems (Ling et al., 2005; Farzan & Brusilovsky, 2006). Designing technical features and seeding social practices to engender continuous contributions from a larger part of the community in recommender system services is an ongoing challenge (Harper et al., 2005; Ling et al., 2005; Farzan & Brusilovsky, 2006). The research is complicated by the fact that motivation is hard to quantify, and while motivation is typically considered to consist of *intrinsic* motivation, i.e. interest in or enjoyment of the task itself within the person, and *extrinsic* motivation, i.e. the task is done to attain something, such as payment, public commendation, or some other benefit, measuring these has turned out to be challenging (Cosley et al., 2005; Neumann, 2007).

Studying the question of motivating users to contribute in recommender systems is roughly divided into two broad branches, 1) motivating users to contribute ratings, and 2) motivating users to provide eWOM, meaning typically ratings and reviews together.

### Why Users Rate Items

User motivations for rating items vary greatly (Ling et al., 2005; Buder & Schwind, 2012). Surveying power users—users that use the system often and rate many items—on MovieLens, Harper et al. (2005) found that in addition to rating items to improve the accuracy of the recommendations they receive, power users rated items for a wide variety of other reasons, e.g. to keep a personal list of movies they had seen, to influence others, and because rating activity itself was felt to be fun and entertaining. Consequently, Harper et al. (2005) feel that in order to motivate users to rate items “it may be more effective to focus on increasing the fun and non-prediction personal benefits of rating through better interfaces for rating and making lists, better interfaces for browsing collections of one’s own ratings, and increased use of games that engage users in the system.”

Herlocker et al. (2004) listed four reasons why users provide ratings:

- *Improving profile* is a motivation that has long been assumed to be the main motivation for rating items; users rate items in order to improve their profile and thus improve the quality of the recommendations they receive.

- *Express self* is a motivation for those users for whom the point is to have a forum to express their opinion and being allowed to contribute is what is important for them (Herlocker et al., 2004). Interviews with power users of MovieLens revealed that for many of them, the motivation was not to improve the recommendations that they received; instead, they rated because it felt good (Herlocker et al., 2004). In fact, for power users with many ratings already in the system, it hardly makes sense to rate more items to improve their profile because the system already has a good profile of them (Konstan & Riedl, 2012).

- *Help others* is a motivation that drives users who believe that their contributions can do good to the community and are happy to help. Many users motivated by the wish to help others are also motivated by self-expression.

- Finally, *Influence others* motivation drives some users whose goal is to explicitly get others to e.g. buy or view particular items or to prevent others from buying or viewing them (Herlocker et al., 2004; Lam, Frankowski, & Riedl, 2006). While these users may be aiming at profit (Lam & Riedl, 2004), many are simply advocates e.g. of some movie genre and rate movies in it highly to push others to see them (Herlocker et al., 2004). Doerr et al. (2012) report that on Digg.com, an email list archive revealed in 2010 showed how "Digg Patriots," a group of users who co-coordinated their actions, successfully gamed the promotion algorithm and managed to push certain stories to the front page in Digg.com. "Digg Patriots" were motivated by anti-liberal agenda and, occasionally, also financial compensation (Halliday, 2010; Doerr et al., 2012).

#### Social Dilemma and Social Loafing: Reasons Why Not to Rate Items

Contributing in recommender systems represents a *social dilemma*, i.e. a situation where it appears rational for individuals not to contribute but it would be better for the collective if all individuals contributed (Cosley et al., 2005; Buder & Schwind, 2012; Cheung & Lee, 2012). Social dilemma is typical to a *public good*, or a shared resource that benefits all members of a group, independent of whether they personally contribute to its provision, and whose use does not diminish its availability (Cheung & Lee, 2012). As rationality, or maximizing self-interest over social interest (Cheung & Lee, 2012), dictates, it has been observed "across a wide range of settings" that people tend to "contribute less than the optimal amount of public goods and consume more than their fair share" (Ling et al., 2005).

In effect, when anyone can access and consume public goods without contributing, as is typical in the case of recommender systems, it is likely to lead to free riding, or social loafing (Ling et al., 2005; Cheung & Lee, 2012). *Social loafing* is a robust phenomenon that describes the tendency for humans to exert less effort on collective tasks than on individual tasks (Karau & Williams, 1993). Contributing, even a simple contribution like rating an item, requires some effort but typically there is no immediate benefit evident for the user (Buder & Schwind, 2012) and especially for an early rater (Burke, 2002).

### Motivating Contributing Ratings and Reducing Social Loafing

Karau and Williams (1993) developed the collective effort model (CEM) to explain social loafing and how people are motivated to contribute to groups (Cosley et al., 2005; Ling et al., 2005). It has been used as a theoretical model to set hypotheses when studying motivations behind contributing ratings, e.g. Cosley et al. (2005), Ling et al. (2005), and Rashid et al. (2006).

Ling et al. (2005) set out to study the propositions of CEM that 1) believing that one's effort is important for the performance of the group[37], 2) believing that one's contributions to the group are identifiable, and 3) liking the group one is working with, should all reduce social loafing and therefore increase contributions (Karau & Williams, 1993).

Using MovieLens community as the test-bed, Ling et al. (2005) found that when people were made to feel that they had unique information to contribute, they did contribute significantly more. On the other hand, those users who received a message emphasizing the benefits for themselves for contributing more actually reduced their number of ratings. In fact, going against the grain of CEM, the users with group goals contributed more than those with individual goals. Also, unlike hypothesized, the group similarity had a negative effect on the number of contributions (usually, people are expected to like better the groups that are similar to themselves).

Ling et al. (2005) also tested the *goal setting theory*, a robust theory of motivation that predicts that assigning people challenging, specific goals motivates them. They found that specific numeric goals did result in more ratings than do-your-best goals. However, the relationship between the size of the goal and the rate of contributing had the shape of an inverse U, suggesting that when the size of the goal became too ambitious, it resulted in demotivation.

Given that their hypothesis received mixed results, Ling et al. (2005) considered that one reason for evidence for some of the hypothesis not to materialize might have been due to a poor implementation—the theory may have been perfectly applicable but the implementation might have failed to apply the design principles appropriately. Psychologists, after all, "have not set themselves the task of designing the optimum configuration of features to achieve a desired outcome" and therefore "provide little guidance about which effect will predominate when a single design feature can have opposite effects" (Ling et al., 2005). In fact, this is

[37] If one's contributions are seen as redundant to what others in the group can or do provide, contributing is not seen as important and there is little incentive to contribute. In contrast, being able to contribute something unique should motivate contributing. (Ling et al., 2005)

something to keep in mind both in designing studies and in looking at their results.

Another factor is that mentioning benefits to self and others might have led to a psychological resistance to emerge; when intrinsically motivated people are provided with extrinsic rewards, the intrinsic motivation often diminishes (Ling et al., 2005; Rashid et al., 2006; Neumann, 2007). In fact, under such conditions, people who are willing to contribute for altruistic or implicit membership reasons may lose their willingness (Neumann, 2007).

Rashid et al. (2006) set out to study the same propositions of CEM as Ling et al. (2005), continuing their work. They wanted to see if making visible the value of the ratings the user provides encourages them to contribute more. In the experiment, they modified the MovieLens interface to indicate how beneficial the rating would be to, depending on the group to which the user belonged, "(1) value to self; (2) value to a small group the user has affinity with; (3) value to a small group the user does not have affinity with; and (4) value to the entire user community." The results did show that explaining the value of the contribution to the members of the community did increase the number of ratings. However, again, when the value was to themselves rather than others, the number of contributions suffered. Also, people rated more movies when the subgroup was similar to them (in terms of liking the same genres) than when the subgroup was dissimilar (liked genres the subject did not). In effect, as predicted by CEM, identifying with the group and liking the group resulted in increasing the number of ratings that subjects made.

Given that showing subjects that they themselves would benefit from rating an item reduced the number of ratings, Rashid et al. (2006) speculated the possibility that subjects perhaps simply did not believe that rating such items would actually enhance the recommendations they received.

While Harper et al. (2007) did not explicitly use CEM as the underlying theory, we discuss their study here, too, as it is thematically related. They studied the impact of social comparisons—used e.g. on Amazon.com that shows a list of top reviewers—with MovieLens as the test-bed by sending subjects an email newsletter that told whether their level of contributing was above, below or about the average in the community. Providing individuals with feedback about their performance or the performance of the group is seen as potentially reducing or eliminating social loafing (Karau & Williams, 1993).

In relation to the control group (who received information about the number of their ratings but no comparison data), all three other groups—including also the group that was given over-performance feedback—

rated more movies during the week following the newsletter. The most motivated group—the one that made most ratings—was the group that had been told that they had rated fewer movies than other members (under-performance feedback). They were also the group to rate more movies in the week following the newsletter than their lifetime per week average, in effect changing their behavior the most. (Harper et al., 2007)

Interestingly, women were more motivated to do something when they were told that they were the same as others (71%) while men were most motivated when they were told that they had made fewer ratings than others (68%). (Harper et al., 2007)

However, the newsletter did not increase overall interest in the site (i.e. users did not visit it more often or click any more links) but rather affected their focus within the system, and so Harper et al. (2007) concluded that social comparison appears to be a potent way to re-direct users' attention. Still, social comparison feedback did have a positive, motivating influence and effectively reduced social loafing, as suggested by Karau and Williams (1993).

These studies show that users tend to respond to social cues. Consequently, in designing recommender system interfaces, it is important to make the community visible and put emphasis on the social aspects of recommender systems, as it can evoke powerful social mechanisms that facilitate rating motivation. While recommender systems are a peer-based technology, they do not offer that much direct peer-to-peer interaction, and so, if not revealed and emphasized, the community remains anonymous and invisible. (Buder & Schwind, 2012)

Also, the value of their contributions needs to be made evident to users to motivate them to contribute; users need to feel that they have something real and unique to contribute to reduce the tendency for social loafing (Karau & Williams 1993; Ling et al., 2005; Rashid et al., 2006; Buder & Schwind, 2012; Konstan & Riedl, 2012). Especially if contributing is the norm in the community, social comparisons can be motivating (Harper et al., 2006; Buder & Schwind, 2012). Moreover, specific, challenging (but not forbiddingly high) goals also help (Ling et al., 2005). Finally, while users are ready to contribute to the community, especially to a sub-group that they like and is similar to themselves, they are not very motivated to contribute to improve the recommendations that they get; i.e. "utility to oneself does not play a major role" (Buder & Schwind, 2012; see also Konstan & Riedl, 2012). Overall, altruism appears to play a greater role in the equation (Hennig-Thurau et al., 2004; Herlocker et al., 2004; Cheung & Lee, 2012) than CEM might lead us to expect.

### Impact of Oversight

*Social facilitation* refers to the effect where "the real or imagined presence of evaluative others results in greater effort on a group task." Sometimes, social facilitation is seen as being related to the social loafing effect in the sense that in facilitation, normative performance feedback is present, and in loafing, it is missing, as feedback can reduce or even eliminate social loafing. (Ling et al., 2005)

Using CEM as the underlying theory, Cosley et al. (2005) set out to study the impact of oversight on the number of contributions, contribution quality, and antisocial behavior on MovieLens. The contribution here was adding movie data to the MovieLens database. The three oversight mechanisms employed were i) *no oversight* (the added movie goes directly to the database), ii) *peer oversight* (the movie goes to the database after one other user has checked the movie information), and iii) *expert oversight* (a movie expert checks the movie information before adding it to the database).

Oversight on MovieLens resulted in less antisocial behavior, a better movie database, and more contributions. Significantly, there was no difference between peer and expert oversight, meaning that both can be used effectively. Also, knowing about oversight had an impact on the quality and quantity of contributions by encouraging good contributors and by discouraging bad ones. (Cosley et al., 2005)

In fact, many online communities have successfully used oversight, e.g. Slashdot.org (to moderate posts), Amazon.com (to rate review helpfulness), and Wikipedia.org (to edit articles and prevent vandalism) (Cosley et al., 2005).

### Motivations for eWOM Contributions

Typically, in studying motivations to contribute in eWOM, the focus is on ratings and reviews that are submitted together as an evaluation. The rating provides (typically) a numeric strength of the evaluation, as the free-text review needs to be read to understand whether it is a recommendation *for* or *against* (Schafer, Konstan, & Riedl, 2001). Users can use ratings to select reviews for reading, e.g. some like to read the negative ones (Leino & Räihä, 2007). Also, ratings make the strength of the recommendation understandable for computers (Schafer, Konstan, and Riedl, 2001). In spite of the importance of eWOM, there has been relatively little research on why consumers engage in it (Cheung & Lee, 2012).

A prominent, oft-quoted study of eWOM motivations by Hennig-Thurau et al. (2004), building on work by Balasubramanian and Mahajan (2001), studied a sample of about 2,000 consumers. They derived four clusters for consumers as far as motivation to engage in eWOM were concerned. As

*Concern for other customers* was the primary motivation for all four segments, the segmentation below is based on secondary motivations:

- *Self-interested helpers* were driven by economic incentives and represent the largest segment with 34% of all respondents.

- *True altruists* were motivated by helping other consumers in addition to helping companies. They represent the second largest segment with 27% of the respondents.

- *Multiple-motive consumers* were motivated by a large number of motives, rating all motives except economic incentives relatively highly. They represented 21% of the respondents.

- The last segment, *Consumer advocates*, had no secondary motive and was simply motivated by concern for other consumers. This, the smallest segment, represented 17% of the respondents. Interestingly, the respondents in this segment had the highest level of formal education of all segments.

Hennig-Thurau et al. (2004) concluded that the primary factors that led to eWOM behavior was concern for other consumers, as well as desire for social interaction, desire for economic incentives, and the potential for enhancing their own self-worth.

Hu, Zhang, and Pavlou (2009) and Hu, Pavlou, and Zhang (2006) posit that the people who provide eWOM do not actually represent the overall consumer public, as there are *two self-selection biases* at work. First, *purchasing bias* results in most eWOM reviews to be positively skewed, as "only people with higher product valuations [are likely to] purchase a product" while people with lower evaluations do not typically purchase the product (Hu, Zhang, & Pavlou, 2009). Consequently, positive reviews are much more common in eWOM (Chevalier & Mayzlin, 2006; Hu, Zhang, & Pavlou, 2009).

Second, it has been found, logically enough, that people engage more in traditional WOM when they feel strongly, be it positively or negatively, about a product than if they are indifferent. Similarly, with eWOM, people who decide to voice their opinions are the ones who are very satisfied with the product or very disgruntled with it. After all, the benefits of writing a review are not obvious and doing so incurs costs, e.g. time and effort. (Hu, Pavlou, & Zhang, 2006)

Consequently, Hu, Zhang, and Pavlou (2009) and Hu, Pavlou, and Zhang (2006) see *bragging* (very positive experience) and *moaning* (very negative experience) as motives for providing reviews. In effect, since mainly consumers with extreme experience—the ones using extreme rating, e.g.

one-star or five-stars on a five-star scale—write reviews and the consumers with moderate views are less likely to write them, the result is *under-reporting bias*, or unrepresentative sample of viewpoints (Talwar, Jurca, & Faltings, 2007; Hu, Zhang, & Pavlou, 2009). The brag or moan model together with *purchasing bias* result in J-shaped distribution typical to product reviews in e-commerce (Hu, Zhang, & Pavlou, 2009).

Hu, Zhang, and Pavlou (2009) discounted the rival explanations, extreme tastes and overconfidence, by conducting an experiment where subjects were asked to rate items that had J-shaped ratings on e-commerce sites. The resulting ratings gave products clearly lower average ratings than on the e-commerce site. Also, the distributions were unimodal, not J-shaped.

Similar to the brag-and-moan model, Hennig-Thurau et al. (2004) suggested the concept of *homeostase utility*, meaning that people desire balance and strive to restore equilibrium if balance is disturbed. In this context, balance can be shaken by strongly a positive or negative consumption experience, leading to, respectively, *expressing positive emotions* or *venting negative feelings*—in effect, to bragging or moaning in terms of Hu, Pavlou, and Zhang (2006).

In effect, the brag or moan model has a good explanatory power on why people engage in eWOM and why e-commerce distributions are typically J-shaped (Hu, Zhang, & Pavlou, 2009). At the same time, it raises the question of whether eWOM actually gives an accurate picture of products (Talwar, Jurca, & Faltings, 2007; Jurca & Faltings, 2009; Hu, Zhang, & Pavlou, 2009).

Cheung and Lee (2012) built on social psychology literature to identify "a number of key motives of consumers' eWOM intention", i.e. antecedents to the intention to engage in eWOM. They tested their model with 203 members of OpenRice.com and found that it explains 69% of the variance based on reputation, sense of belonging, and enjoyment of helping other consumers. Sense of belonging had the strongest impact on eWOM intention. Also, enjoyment of helping others had a strong impact. Reputation, while impacting the equation, was only a marginally significant factor. Consequently, Cheung and Lee (2012) see it as important to enhance users' sense of belonging to an online opinion platform, to use reputation-tracking mechanisms to give recognition to good contributors, and to give feedback to users whose contributions have helped others, to motivate contributing eWOM.

On the other hand, reciprocity, moral obligation, and knowledge self-efficacy[38] did not have significant relationship with eWOM intention. As

[38] S*elf-efficacy* is our personal judgment of our ability to do what is required to accomplish something (Cheung & Lee, 2012).

OpenRice.com is an informal community, it may be that, while members can have a sense of belonging, they do not typically develop a sense of duty or obligation towards it. Also, since first-timers are invited to share their experiences and there is no formal criteria for contributions to meet, knowledge self-efficacy may have ended up playing a smaller role here than in some other settings. (Cheung & Lee, 2012)

**In summary**, the results of Hennig-Thurau et al. (2004) and Cheung and Lee (2012) are largely in agreement and underline *altruistic* motivations, helping other consumers, and *collectivist* motivations, sense of belonging and desire for social interaction, in addition to somewhat less important *egoistic* motivations, reputation and enhancing self-worth, as the main motivations for contributing eWOM. The results of Hu, Zhang, and Pavlou (2009) and Hu, Pavlou, and Zhang (2006), that contributing is driven by extremely positive (brag) or negative (moan) consumption experiences, is very similar to Hennig-Thurau et al.'s (2004) concept of homeostase utility of venting negative feelings and expressing positive emotions. People contribute when they have strong feelings (Hu, Pavlou, & Zhang, 2006), which can lead to unrepresentative sample of views in eWOM and raises questions concerning the accuracy of the picture that eWOM provides of products (Talwar, Jurca, & Faltings, 2007; Jurca & Faltings, 2009; Hu, Zhang, & Pavlou, 2009). Consequently, one of the major questions in eWOM is how to motivate those consumers who have moderate views to contribute their views.

# 8 Evaluating Recommender Systems

Evaluating recommender systems is innately difficult and it has been approached in many, often dissimilar ways (Herlocker et al., 2004). In addition to algorithms tending to perform better or worse depending on data set characteristics, the goals and purposes for which recommender systems are developed and evaluated differ, meaning that no one evaluation technique is going to suit them all (Herlocker et al., 2004; Schafer et al., 2007; Redpath et al., 2010).

In general, perspectives to evaluating recommender systems can be divided into system-centric and user-centric, although the two approaches can also be combined. *System-centric* approaches evaluate recommenders against a pre-built or pre-collected dataset of user preferences using such quality measures as precision and recall without users interacting with the system during the test; the user opinions on the items have been gathered beforehand and testing is done against values that are withheld from the dataset available to the recommender. (Herlocker et al., 2004; Cremonesi, Garzotto, & Turrin, 2013)

In contrast, *user-centric* approaches have users interact with a running recommender—or recommenders if two or more variations are being compared—and the data is collected during or based on the interaction; users are either asked, e.g. through interviews or surveys, or their behavior is observed during the use or their interactions are recorded and then analyzed, e.g. for click through or conversion rates (Herlocker et al., 2004; Cremonesi, Garzotto, & Turrin, 2013). User experiments can be controlled or field studies, i.e. in the wild (Herlocker et al., 2004).

On a side note, Herlocker et al. (2004) called these two approaches *offline analysis* and *live user experiment* but the idea is the same. There are also other ways of categorizing approaches to evaluating recommender systems, e.g. Konstan (2008) talks about metrics, research measures (e.g. user satisfaction), and industry outcomes (e.g. add-on sales and click-through rates), but, by and large, system-centric and user-centric categories cover them; e.g. system-centricity covers Konstan's (2008) metrics and user-centricity covers research measures and industry outcomes.

Despite Shardanand and Maes already underlining in 1995 that the human response is more important than "numbers" and Swearingen and Sinha in 2001 pointing out that other factors than algorithmic aspects are also important, evaluating recommender systems has largely focused on using various (statistical) metrics to evaluate algorithms offline (Herlocker et al., 2004; Cremonesi et al., 2011; Knijnenburg, Willemsen, & Kobsa, 2011; Buder & Schwind, 2012; Knijnenburg et al., 2012; Pu, Chen, & Hu, 2012), as the implicit assumption has been that better algorithms result automatically in a better user experience (Knijnenburg et al., 2012).

More recently, however, there has been an increasing emphasis on user-centric methods, as it has become clear that system-centric measures ignore many crucial aspects that determine recommender adoption and perceived usefulness (Herlocker et al., 2004; McNee, Riedl, & Konstan, 2006a; Cremonesi et al., 2011; Knijnenburg et al., 2012; Pu, Chen, & Hu, 2012). In fact, McNee, Riedl, and Konstan (2006a) state that the narrow focus on algorithms has not only been misguided but in fact "detrimental to the field." Consequently, there is a growing consensus that the bottom-line measures for recommender systems should be user-centric, i.e. such measures as user satisfaction and user experience (Herlocker et al., 2004; Redpath et al., 2010; Cremonesi et al., 2011; Buder & Schwind, 2012; Knijnenburg, 2012; Knijnenburg et al., 2012; Pu, Chen, & Hu, 2012). That is not to say that work on algorithms is not important or necessary but that evaluating algorithms offline cannot be seen as the be-all-end-all of evaluating recommender systems (Swearingen & Sinha, 2001; McNee, Riedl, & Konstan, 2006a; Cremonesi, Garzotto, & Turrin, 2013).

Underlining the importance of user-centric approach, using the perspective of user experience to look at recommender systems can lead to counterintuitive results (Konstan & Riedl, 2012). For example, McNee et al. (2003) showed how a slower initial rating interface that required more user effort but gave users more control led to higher user retention and improved user trust and satisfaction, even though the bigger rating effort did not improve actual prediction quality. Such counterintuitive findings underscore the importance of user-centric evaluation approaches, as we otherwise end up designing systems that are not satisfactory but we have no means to understand why they are not satisfactory.

Overall, perhaps surprisingly, there have not been many studies that have compared the results of system-centric evaluations to user-centric evaluations in live experiments (Knijnenburg et al., 2012). However, the ones that have been conducted have largely suggested that there is little correlation between algorithmic accuracy and user experience—algorithmically "best" systems have not been the best systems when measured using user-centric methods (Cremonesi et al., 2011; Knijnenburg et al., 2012; Pu, Chen, & Hu, 2012; Cremonesi, Garzotto, & Turrin, 2012a; Cremonesi, Garzotto, & Turrin, 2013). Objective metrics, i.e. statistical accuracy metrics, in effect, do not consistently predict well the perceived quality of recommender systems (Cremonesi et al., 2011).

For example, Cremonesi et al. (2011) found in an experiment in the domain of movie recommending that "simple non-personalized TopPop recommendations are better perceived by the users with respect to other more sophisticated and personalized recommender algorithms, although users are aware of the low utility of such recommendations." Consequently, some researchers have suggested using more user-centric concepts like perceived quality or perceived accuracy that may influence more behavioral intentions (Cremonesi et al., 2011; Pu, Chen, & Hu, 2012).

One of the rare exceptions to this trend is a recent study by Cremonesi, Garzotto, and Turrin (2013) in the domain of tourism, specifically booking a hotel online, that found that system-centric measures were, in fact, consistent with user-centric measures. The authors point out that the results may be due to peculiarities of the domain where the novelty or serendipity of the recommendations is not necessarily as important an aspect as in many other domains and where even optimizing users may be forced to make sub-optimal decision e.g. if the "best" hotel is fully booked. They conclude that "the relationship between the two kinds of metrics may depend on the business sector, is more complex that we may expect, and is a challenging issues [sic] that deserves further research."

There are simple reasons why such objective statistical metrics as various error and accuracy metrics have traditionally been used in evaluating recommender systems without involving users directly. After all, offline evaluations, or systems-centric evaluations, are quick and economical to conduct even on a large scale even if several datasets and algorithms are involved (Herlocker et al., 2004; Cremonesi, Garzotto, & Turrin, 2013). Also, such data is unequivocal in the way that e.g. behavioral data is not (Knijnenburg et al., 2012). As a result, many issues that are important to users but complex to operationalize have been simply neglected or ignored (Cremonesi et al., 2011). There are also other weaknesses to offline evaluations resulting e.g. from data sparsity that "limits the set of items that can be evaluated" (Herlocker et al., 2004).

Moreover, it has been argued that evaluating recommenders by such measures as prediction accuracy (e.g. mean absolute error or (root of) mean square error), decision support accuracy (e.g. reversals or receiver operating characteristics), rank recommendation metrics (e.g. top-n measures or half-life), or item-set coverage actually measure wrong things as far as actual users of recommender systems are concerned (McNee, Riedl, & Konstan, 2006a; Konstan, 2008; Bobadilla et al., 2013). Such measures do not help users find things that they do not know about but would like, value, or use, and instead reward recommender algorithms for recommending things that users actually already know about (McNee, Riedl, & Konstan, 2006a; Konstan, 2008; Konstan & Riedl, 2012). For example, a recommender system that recommends *The White Album* by The Beatles to fans of the band would be extremely accurate but practically useless (Herlocker et al., 2004; McNee, Riedl, & Konstan, 2006a). The same goes for a travel recommender that would recommend only places that we have already visited instead of finding us new places to visit (McNee, Riedl, & Konstan, 2006a). However, this is exactly how accuracy metrics test recommenders (McNee, Riedl, & Konstan, 2006a; Konstan, 2008; Konstan & Riedl, 2012).

Moreover, this kind of evaluating can lead to the rich-getting-richer phenomenon where popular, already-rated items get recommended again and again while unpopular items and items with limited historical data do not get recommended (Cremonesi et al., 2011; Konstan & Riedl, 2012). Recommenders may lead to winner-takes-all situations and recommenders may, in fact, end up working against their purpose of helping users find new, salient items, resulting in homogenization (Konstan & Riedl, 2012).

Perhaps the greatest problem of offline evaluations is related to user goals and preferences that cannot be measured offline (Herlocker et al., 2004; McNee, Riedl, & Konstan, 2006a; Redpath et al., 2010; Knijnenburg, Willemsen, & Kobsa, 2011; Konstan & Riedl, 2012). It is user satisfaction (that is affected by many other factors than accuracy and does not even necessarily correlate with accuracy) that is the most important aspect to evaluate (Herlocker et al., 2004; McNee, Riedl, & Konstan, 2006a; Knijnenburg, Willemsen, & Kobsa, 2011; Knijnenburg, 2012; Pu, Chen, & Hu, 2012). Consequently, there is a growing consensus that evaluating recommender systems should involve actual users and actual use (Schafer, Konstan, & Riedl, 2001; Konstan, 2008; Redpath et al., 2010; Knijnenburg, 2012).

In fact, e-commerce has long recognized this problem and has striven to adapt its algorithms to serve user experience as well as sales and marketing efforts, resulting in a greater emphasis on serendipitous recommendations right from the start (Konstan & Riedl, 2012). In effect, recommender systems need to help users make better decisions, and accuracy alone, especially as measured today, is not enough (Schafer, Konstan, & Riedl, 2001; Knijnenburg, Willemsen, & Kobsa, 2011). In fact, today there is a growing interest to generate algorithms that serve users better "even at the expense of accuracy and precision" (Bobadilla et al., 2013).

In spite of the field knowing that user-centric approaches to evaluating recommender systems are crucially important, they have not been employed anywhere near to the same degree as system-centric approaches (Cremonesi, Garzotto, & Turrin, 2012a; Knijnenburg et al., 2012). The reason is that, unlike system-centric approaches, user-centric approaches are complex, difficult, expensive, and resource-demanding (Kumar & Benbasat, 2006; Redpath et al., 2010; Cremonesi et al., 2011; Knijnenburg, Willemsen, & Kobsa, 2011; Buder & Schwind, 2012; Konstan & Riedl, 2012; Cremonesi, Garzotto, & Turrin, 2013). As Konstan and Riedl (2012) point out, "measuring user experience requires developing a system, including both algorithms and user interface, and carrying out field studies with long-term users of the system—the only reliable way of measuring behavior in a natural context."

In effect, conducting human-centered research on recommender systems requires finding or maintaining user communities where recommenders are employed, meaning that there is a need to i) develop systems dedicated to experimental use; ii) collaborate with operators of live systems; or iii) developing and maintaining research systems and user communities (Konstan & Riedl, 2012). The challenges involved are compounded by related factors, e.g. developing a large and diverse enough user group or community, and designing and implementing recommender systems (Redpath et al., 2010; Buder & Schwind, 2012; Cremonesi, Garzotto, & Turrin, 2012a). At the same time, most organizations keep their data private for both privacy and competitive reasons and, typically, only larger companies or research labs have expertise and resources to develop and run such systems (Schafer et al., 2007). Moreover, as things stand, user experience in recommender systems remains "an ill-defined concept" that "lacks well-developed assessment methods and metrics" (Knijnenburg et al., 2012). As if these challenges were not daunting enough, in addition to many issues being complex to operationalize, the high number of variables that needs to be controlled results in intrinsic complexity (Cremonesi et al., 2011).

Consequently, in approaching evaluating recommender systems in a user-centric fashion, there is a need to use both explicit and implicit data, and to use both observation and such methods as interviewing or questionnaires (Herlocker et al., 2004; Knijnenburg et al., 2012). Implicit data can refer to both observing users and to log data of actual use while explicit data refers to explicitly soliciting user viewpoints e.g. by interviews or questionnaires (Herlocker et al., 2004; Knijnenburg et al., 2012). While having both types of data helps us to combat the say-do problem (Leino & Räihä 2007; Leino, 2013), behavioral data is not always a good indicator of subjective experience and interpreting it can be troublesome, and having explicit data from users can help solve such problems (Knijnenburg et al., 2012). In effect, the two types of data can also be used for triangulating for the best results (Herlocker et al., 2004; Knijnenburg et al., 2012).

Finally, the interface of the system and the decisions made in designing it can dramatically impact the user behavior in and experience of the system (Lerman, 2007; Knijnenburg et al., 2012), as such complex systems result in emergent behavior that makes predicting behavior and experience challenging (Lerman, 2007). In summary, it is easy to agree with Herlocker et al. (2004) that "[e]ffective and meaningful evaluation of recommender systems is challenging."

Below we discuss user-centric evaluation and user experience in recommender systems in further detail but leave system-centric measures outside of the discussion, as all the publications in this dissertation have approached recommenders from user-centric viewpoint[39]. First, however, we look briefly at metrics that go beyond accuracy.

## 8.1 BEYOND ACCURACY

As accuracy is not the only or perhaps not even the most important criteria of interest in all instances, several other evaluation criteria have been suggested and explored, leading to new notions of quality that paint a fuller picture of the actual performance of recommenders (Herlocker et al., 2004; Schafer et al., 2007; Konstan & Riedl, 2012). In fact, we have already discussed some of these newer criteria, e.g. diversity, serendipity, explanations, and familiarity, above (Section 5.11), and, consequently, as these metrics are not in the focus of this dissertation, we will only discuss some new ones briefly here.

[39] Readers interested in statistical/mathematical details of system-centric evaluating of recommenders are directed to Herlocker et al. (2004), as this is the seminal paper in the field. For further reading, works by Cremonesi with various co-authors, e.g. Cremonesi et al. (2008) and Cremonesi et al. (2011), are perhaps more detailed but less comprehensive in their explanations, as they only discuss system-centric measures to the degree that is relevant to their current work. In addition, Bobadilla et al. (2013) provide a good section of system-centric evaluation measures in their article.

### Lift and Hit

Business application of recommenders brought out new vocabulary and metrics for evaluating recommenders as practical concerns forced researchers and designers "to think more broadly about both the evaluation and the design of recommender systems and interfaces" (Konstan & Riedl, 2012). Such metrics as lift and hit rate were seen as providing better evaluation of how a recommender was performing (Konstan & Riedl, 2012). *Lift factor/rate* refers to the increase of response, e.g. sales, caused by the introduction of the recommender system (in relation to the baseline situation of not having the recommender system), i.e. the resulting increase in sales, while *hit rate*, or *conversion rate*, refers to the percentage of recommendations that are converted into action, e.g. sales, i.e. the actual frequency with which recommended items are selected (Cremonesi, Garzotto, & Turrin, 2012a; Konstan & Riedl, 2012). In other words, recommenders are evaluated in terms of the *behavioral impact* they have rather than in terms of theoretical accuracy of the algorithm. This is natural in the context of e-commerce where the goal is to sell, and therefore it makes sense to evaluate recommenders in terms of sales. Miniscule improvements in the accuracy that are not perceptible to users and that cause no behavioral impact are hardy meaningful in this context (Swearingen & Sinha, 2001; Herlocker et al., 2004; Konstan & Riedl, 2012).

### Avoiding Bad Mistakes

Avoiding bad mistakes is also another metric that has a strong impact on user trust (Konstan & Riedl, 2012). In 1995, Shardanand and Maes had already recognized that errors in the cases of very high (highly liked items) and very low (not at all liked items) ratings would easily undermine trust, as these are the cases where the user has a strong feeling about the item. In fact, Herlocker et al. (1999) introduced receiver operating characteristic curve (ROC curve) that only penalizes cases where a prediction would result in a high-salience item to be missed or low-salience item to be consumed for evaluating prediction algorithms (see also Konstan & Riedl, 2012).

### User Control

Users should feel that they control their interactions with recommenders (Xiao & Benbasat, 2007; Konstan & Riedl, 2012). Giving users more control may even involve increasing the user effort required and result in recommendations being *less* accurate by objective measures while still resulting in increased user satisfaction (Konstan & Riedl, 2012). However, it appears that the general trusting propensity of the user plays a role here, too (Knijnenburg, Reijmer, & Willemsen, 2011). Distrusting users typically want to have more control over the system, although explanations about how the system works can lower this desire (Knijnenburg, Reijmer, & Willemsen, 2011). In addition, giving users a feeling of control entails involving users through feedback: While the response times should be kept at a minimum, when users are required to wait, recommenders should provide them with information about the progress (Xiao & Benbasat, 2007). Recommender design needs to balance giving the user a feeling of control and serving them effectively and unobtrusively (Konstan & Riedl, 2012).

#### Robustness

Given all the unscrupulous "nuking" and "pushing", i.e. shilling, that is going on, it is no wonder that recommender system robustness against (outside) manipulations is an area of intense research efforts (Herlocker et al., 2004; Konstan & Riedl, 2012). However, recommender systems continue to be vulnerable to various types of attacks and such attacks are difficult to detect, even if their impact on popular items in large communities is lower due to the countervailing data (Konstan & Riedl, 2012; Bobadilla et al., 2013). There are no good metrics to measure recommender robustness yet (Herlocker et al., 2004; Konstan & Riedl, 2012; Bobadilla et al., 2013).

#### Coverage

Coverage typically means the percentage of items in the system for which predictions can be made (Herlocker et al., 2004; Schafer et al., 2007; Su & Khoshgoftaar, 2009) although there are also variations, e.g. "the percentage of items that have the potential of being recommended to users, as performance optimizations in recommendations may prevent certain items from ever being recommended" (Schafer et al., 2007). High coverage is especially important for *Find all good items* task, as items without ratings or other preference data cannot be found, and for *Annotate in context* task, as annotating is simply not possible without data (Herlocker et al., 2004). Of course, low coverage is problematic in any scenario, as the recommender system decision quality is inherently low if the system is unable to consider a large number of available items (Herlocker et al., 2004). Nevertheless, no general coverage metric exists (Herlocker et al., 2004).

#### Learning Rate

Learning rate refers to the amount of data necessary for an algorithmic recommender system to become effective in predicting user tastes (Herlocker et al., 2004; Schafer et al., 2007). Three learning rates have been considered: overall learning rate, i.e. the recommendation quality as a function of the overall number of ratings (or users) in the system; per item learning rate, i.e. "the quality of predictions for an item as a function of the number of ratings available for that item"; and per user learning rate, i.e. "the quality of the recommendations for a user as a function of the number of ratings that [the] user has contributed" (Herlocker et al., 2004; see also Schafer et al., 2007). The rate at which different algorithms reach an acceptable recommendation quality can vary greatly from only a few data points to needing a very high number of data points (Herlocker et al., 2004). Learning rates are by necessity non-linear and asymptotic, as quality cannot be endlessly improved (Herlocker et al., 2004). Consequently, learning rates of different algorithms are typically compared by graphing the quality versus the number of ratings (quality is typically represented by accuracy) (Herlocker et al., 2004).

#### Confidence

Confidence refers to the recommender system's ability to evaluate the likely quality of its recommendations (Schafer et al., 2007). In effect, the user needs to decide how to interpret the recommendations a system provides along two potentially conflicting dimensions: The *strength* of the recommendation, i.e. how much does the recommender think the user will like the item, and the *confidence* of the recommendation, i.e. how sure the recommender is that its suggestion is accurate (Herlocker et al., 2004). In practice, very high predictions, e.g. that the user will like an item at five stars on a five-star scale, tend to be based on a small amount of data (Herlocker et al., 2004). Consequently, to help users make good decisions based on the recommendations, the recommender system should also help them assess the suggestions based on confidence and not only strength. In e-commerce, the typical solution is not to show recommendations when the dataset on which the strength is based is too small; the goal is not to look stupid—and lose user trust—and to provide recommendations that users can trust, i.e. have high enough confidence (Herlocker et al., 2004). When confidence can be computed, it can be displayed to help users judge the suggestions more effectively (Herlocker et al., 2004; Schafer et al., 2007). However, measuring the quality of confidence is challenging, as confidence is "a complex multidimensional phenomenon that does not lend itself to simple one-dimensional metrics" (Herlocker et al., 2004).

**In summary**, accuracy is only one criteria or measure of interest and, at least in some cases, it might not even be the most important one (Schafer et al., 2007). There is a movement in the field to move beyond accuracy, to explore and develop metrics for many other aspects of recommender systems that would allow us to evaluate recommenders better and to provide a better user experience (Herlocker et al., 2004; McNee, Riedl, & Konstan, 2006a; Konstan & Riedl, 2012). Designers of recommender systems need to carefully consider how they approach evaluating recommender systems to make sure that they actually measure the value they provide to the users (Konstan & Riedl, 2012).

## 8.2 USER-CENTRIC EVALUATING AND USER EXPERIENCE

Perhaps the greatest shortcoming of system-based metrics is their narrow focus on one technical aspect of recommender systems. While metrics that go beyond accuracy, such as diversity, coverage, and confidence, take a wider approach to cover and better evaluate what makes recommenders useful for real-life users, they still only cover certain aspects that influence user satisfaction and experience. Moreover, such measures as serendipity and perceived accuracy, or perceived usefulness, for that matter, cannot be measured without involving users. Finally, there are no metrics to measuring community opinion recommenders; users need to be involved in evaluating such measures.

Consequently, it is increasingly being argued that the maxim "[u]ser experience is everything" (Williams, 2005) also applies to recommender systems, as the purpose of recommenders is to help users and to provide value to users (McNee, Riedl, & Konstan, 2006a). Therefore, recommenders need to be evaluated in terms of whether or not they help users meet their goals, the benefit they provide their users with, and what kind of user experience they provide (Herlocker et al., 2004; McNee, Riedl, & Konstan, 2006a). Furthermore, next to accuracy and helpfulness, recommenders must also be made fun and pleasurable to use (McNee, Riedl, & Konstan, 2006a; Knijnenburg et al., 2012). Consequently, many see user-centric approaches to evaluating recommenders, i.e. involving users in evaluating recommenders, as essential (e.g. Herlocker et al., 2004; McNee, Riedl, & Konstan, 2006a, Redpath et al., 2010; Pu, Chen, & Hu, 2011; Konstan & Riedl, 2012; Cremonesi, Garzotto, & Turrin, 2012a; Pu, Chen, & Hu, 2012).

However, as discussed, the emphasis on user-centric approaches does not mean to claim that accuracy metrics should not be used, only that they are not enough alone (McNee, Riedl, & Konstan, 2006a).

#### User Experience

At the core of the issue is the concept of *user experience*. The problem with the concept is that in recommender systems it remains an "ill-defined concept" that "lacks well-developed assessment methods and metrics" (Knijnenburg et al., 2012; see also Cremonesi, Garzotto, & Turrin, 2012a; Pu, Chen, & Hu, 2011).

However, we do have some recent, if perhaps tentative, definitions at our disposal. Konstan and Riedl (2012) define user experience as "the delivery of the recommendations to the user and the interaction of the user with those recommendations" while Knijnenburg et al. (2012) define it as "the user's evaluation of the system (perceived system effectiveness and fun), system usage (usage effort and choice difficulty), and outcome of system usage (satisfaction with the chosen items)." Konstan and Riedl (2012) see their own definition as "grounded in specific recommender systems and their evaluations" and as standing "in contrast to Knijnenburg et al. (2012) which approaches user experience from more of an experience-model and social-experimental approach."

Neither definition places any great emphasis on the preference elicitation phase that typically precedes the delivery of recommendations and colors the subsequent experience (Gretzel & Fesenmaier, 2006; Xiao & Benbasat, 2007; Pommeranz, Wiggers, & Jonker, 2011); the definition by Konstan and Riedl (2012), in fact, cuts it out entirely, starting with the delivery of recommendations, while the definition by Knijnenburg et al. (2012) does encompass it but appears to emphasizes more the experience related to the outcome.

Xiao and Benbasat (2007) see recommender systems as consisting of three major components: Input (elicitation of user preferences explicitly or implicitly), process (generation of recommendations), and Output (presenting recommendations to users). Accuracy metrics have largely focused on the process and the two other components have not been considered widely (Xiao & Benbasat, 2007). However, the two other components cover many aspects, e.g. provider/source credibility, and product- and user-related factors that influence user experience (Xiao & Benbasat, 2007; Knijnenburg et al., 2012). In effect, Knijnenburg et al. (2012) argue that user experience does not only depend on the recommender system, as it is also affected by personal characteristics of the user and the situational characteristics of the situation in which the use takes place. Therefore, output, or presenting recommendations, cannot be the end point of user experience; we still need to evaluate if the need(s) of the user were met (Herlocker et al., 2004; McNee, Riedl, & Konstan, 2006a)—we still need to consider the outcome of the interaction with the system, as the definition by Knijnenburg et al. (2012) points out.

Similar to Xiao and Benbasat (2007), Pu, Chen, and Hu (2012), in their generic interaction model, see the interaction between users and recommenders as taking place in three steps: Elicit preference (explicitly, implicitly, or any combination of the two; users browse, search, purchase, rate etc.), display recommendations (presenting recommendations; users view recommendations), and revise preference (revising preferences based on user feedback; users give feedback by rating, critiquing, selecting recommended items etc.). The difference to the model by Xiao and Benbasat (2007) is that, in the last step, user reactions to the output are used as feedback to the system; the ease of generating additional recommendations has been shown to increase perceived user control, leading to increased trust and satisfaction (Xiao & Benbasat, 2007; Pu, Chen, & Hu, 2011).

The central point is that the concept of user experience extends the evaluating of recommender systems outside of the process stage to cover the whole of the user interaction with the recommender system. In addition, user experience, as defined in ISO 9241-210—"person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service"—also includes anticipated use, and there is hardly a reason not to include it also when talking about user experience in recommender systems.

In fact, as user experience in the field of recommender systems remains an "ill-defined concept" where the various aspects that affect it have not been integrated into a model of overall user experience (Knijnenburg et al., 2012), indeed have not even been investigated systematically (Cremonesi, Garzotto, & Turrin, 2012a), there continues to be a need to define user experience in recommender systems by the constituents of an effective and satisfying recommender system (Pu, Chen, & Hu, 2011) so that studying it systematically is possible.

Since the concept of user experience has been considered and defined in general terms for information systems in general, it behooves to look at those definitions as a starting point. As mentioned, there is an ISO definition of user experience that brings anticipation into picture, something that the definitions within the field of recommender systems have not emphasized. In addition, Bevan (2009) suggests that "perceptions and responses" mentioned in the ISO 9241-210 definition can be likened to the notion of *satisfaction* that is used in the ISO 9241-11, i.e. the ISO definition of usability. Likewise, when discussing the user-centric approach to evaluating recommender systems, many researchers use the concept of user satisfaction analogously or semi-analogously to user experience (e.g. Herlocker et al., 2004; McNee, Riedl, & Konstan, 2006a; Schafer et al., 2007; Xiao & Benbasat, 2007; Konstan & Riedl, 2012).

Moreover, the concept of user experience has placed emphasis on aesthetics and engagement in addition to underlining the importance of emotion and affect (Day et al., 2010). Similarly, the importance of interface and presentation to the overall experience of recommender systems is increasingly recognized (e.g. Gretzel & Fesenmaier, 2006; Knijnenburg et al., 2012; Pommeranz et al., 2012). Also, significantly, the concept of user experience emphasizes holistic aspects of evaluating the interaction with information technology rather than focus on some aspects (Hassenzahl & Tractinsky, 2006; Day et al., 2010). Hassenzahl and Tractinsky (2006) conclude that user experience is about "designing for pleasure, rather than absence of pain." In other words, technology—and recommender systems by inference—should be pleasurable and fun to use, and not only accurate and helpful, just as McNee, Riedl, and Konstan (2006a) have argued within the field of recommender systems. Based on the above considerations, Day et al. (2010) suggested the following definition for user experience:

> "User experience is a person's perceptions and responses that result from the use and/or anticipated use of a product, system or service, where these perceptions and responses can

> be characterised in terms of effectiveness, efficiency and satisfaction in achieving specified goals, and also in terms of an engagement with the product, system or service."

Importantly, Day et al. (2010) see *engagement* as encompassing persuasion, something that is central in recommender systems (Gretzel & Fesenmaier, 2006; Knijnenburg, Reijmer, & Willemsen, 2011; Buder & Schwind, 2012; Cremonesi, Garzotto, & Turrin, 2012a).

Characterizing "perceptions and responses" in terms of "effectiveness, efficiency, and satisfaction" makes it feasible to measure user experience (Day et al., 2010)—and without measurability the concept would be rather worthless for evaluating recommender systems. Consequently, based on Day et al. (2010), we suggest the following definition for user experience in recommender systems:

> *A person's perceptions and responses that result from the use and/or anticipated use of a recommender system, where these perceptions and responses can be characterized in terms of effectiveness, efficiency, and satisfaction in achieving specified goals, and also in terms of an engagement with the recommender system.*

This definition captures the need to see the interaction as a whole, therefore also encompassing anticipation and expectations that color and frame the actual interaction. Also, it incorporates emotional, aesthetic, and persuasive aspects. Importantly, it also incorporates the user task and goal, as recommender systems are useful only to the extent that they allow users to complete their tasks and reach their goals (Herlocker et al., 2004; Redpath et al., 2010). By including specific user tasks, the definition also encompasses situational characteristics, the situation in which the interaction takes place, and by focusing on the user's perceptions and responses also encompasses personal characteristics of the user, i.e. factors that impact user experience but that cannot be influenced by the recommender system, at least not directly (Knijnenburg et al., 2012). For example, users with high product class knowledge tend to trust more and be more satisfied with content filtering recommender systems than with collaborative filtering systems, while the reverse was true for users with low knowledge (Xiao & Benbasat, 2007), and users with a high level of domain knowledge tend to find recommender systems less useful and more difficult to use than those with a novice level of domain knowledge (Knijnenburg et al., 2012).

Finally, what aspects constitute a good experience may change over time; e.g. a new user needs to establish trust in the system and needs more familiar recommendations, while later on the same user, since they have by now gained more experience of the system and established trust, needs more serendipitous recommendations to perceive the system as useful and satisfactory (Herlocker et al., 2004; McNee, Riedl, & Konstan, 2006a).

### Conducting User Evaluations and Experiments

Given the importance of user experience to evaluating recommender systems, there remains the question of how to evaluate user reaction to and experience of a recommender system. Based on a literature review, Herlocker et al. (2004) discuss four dimensions for user evaluations: Explicit vs. implicit, laboratory vs. field, outcome vs. process, and short-term vs. long-term.

***Explicit* (asking) *vs. implicit* (observing)**: There are basically two ways to get user-centric information, explicitly, i.e. by asking users (interviews, questionnaires, and similar methods), and implicitly, i.e. by observing user behavior (Herlocker et al., 2004). Observation is typically done by logging user behavior and then subjecting the data to analysis (Herlocker et al., 2004). However, observation can also be done by observing the user while they are using the systems by being present at the use situation (e.g. Leino & Räihä, 2007), by taping the use situation, or by screen capturing during the use, to name a few alternatives.

Whenever possible, the implicit and explicit collecting of data should be combined (Herlocker et al., 2004; Knijnenburg, Willemsen, & Kobsa, 2011; Knijnenburg et al., 2012). First, user preference and performance may differ; e.g. the user may prefer one system even if there is no difference in the performance (Herlocker et al., 2004). In effect, behavioral data is not always a good indicator of the subjective experience, and, moreover, interpreting it can be challenging (Knijnenburg, Willemsen, & Kobsa, 2011; Knijnenburg et al., 2012). Second, using both approaches together allows for triangulating (Herlocker et al., 2004; Knijnenburg, Willemsen, & Kobsa, 2011; Knijnenburg et al., 2012). Moreover, using only *say* data, or explicit data, without *do* data, or implicit data, there is a risk of the say-do problem (Kobsa, 2007), as “What people say, what people do, and what they say they do are entirely different things”, as Margaret Mead said[40].

In general, qualitative data (interviews, open textual questionnaire questions, observation by looking at use, etc.) is more explanatory but less generalizable and challenging if not impossible to validate statistically, while quantitative data (use statistics, evaluative statements that use an interval scale in questionnaires etc.) allows for statistical validation but typically requires theoretical hypotheses about theoretical constructs and their relations (Knijnenburg et al., 2012).

***Laboratory studies vs. field studies***: Laboratory studies are good for testing hypotheses and are therefore useful in investigating specific issues.

[40] https://en.wikiquote.org/wiki/Margaret_Mead

Field studies, on the other hand, reveal what users actually do in actual use contexts, laying bare common use and use patterns, problems and unmet needs, and also issues that the runners of the study had not thought of beforehand. (Herlocker et al., 2004)

Significantly, when the task does not really involve participants, often the case in laboratory experiments, especially when students are used as proxies for actual users, or the task is trivial, or participants are not motivated, participants tend to use simple heuristics, e.g. affect, to make decisions (Sillence & Briggs, 2007). In contrast, in high-involvement situations, when participants have something to lose or are deeply engaged, they tend to use "cognitive analytical processing" (Sillence & Briggs, 2007).

***Outcome vs. process***: In addition to there being a need to clearly define user tasks that the system is meant to support, appropriate metrics must also be developed relative to the particular task to determine what constitutes a successful outcome. Simply measuring if the goal is achieved is not enough, as systems may differ in how efficiently they allow users to reach their goal; for example, amount of time and effort required should be measured "to make sure that the cost of a successful outcome does not outweigh the benefit." (Herlocker et al., 2004)

***Short-term vs. long-term***: Some issues may not become evident in short-term studies, in particular short-term laboratory studies. True adoption of a system, for example, can only be ascertained with long-term studies. We must consider if the question can be answered with a short-term study or if a long-term set-up is the only possible solution. (Herlocker et al., 2004)

# 9 Recommender Domains Relevant to the Publications

While recommender systems are perhaps best known for being used on e-commerce sites, such as Amazon.com, in fact the early research into them started in domains such as newsgroups (Tapestry and GroupLens) (Goldberg et al., 1992; Konstan et al., 1997), music (Ringo) (Shardanand & Maes, 1995), and videos (Bell-core's Video Recommender) (Hill et al., 1995). In effect, as information overload is a growing concern and users are in need of decision-helping aids in a wide range of domains, recommender systems have spread far and wide, both online and offline (Lam, Frankowski, & Riedl, 2006; Konstan & Riedl, 2012; Said et al., 2012).

As covering all the different domains where recommender systems are currently being employed would be a huge undertaking and beyond the scope of this dissertation, we only discuss the three domains connected to the publications that constitute this dissertation, namely e-commerce, news-recommending, and e-learning. Moreover, as the domain of e-commerce is a huge domain in and of itself and has already been widely discussed in this dissertation, as much defining research work has focused on it, here we only very briefly look at recommenders in e-commerce through Amazon.com since it is the subject of the research of the publications focusing on e-commerce and an early and defining pioneer in the field whose approach has been widely imitated (Chalmers et al., 2004; Kumar & Benbasat, 2006; Pathak et al., 2010) and that "is the clear winner in shopper preference concerning recommender systems" (Ozok, Fan, & Norcio, 2010).

## 9.1 E-COMMERCE

Recommender systems were an integral part of such early e-commerce retailers as Amazon.com and CDNOW.com (later acquired by Amazon), as there was a clear need to help people to find salient items and to decide which ones to consume by providing suggestions and such information as community opinion and critique about the items (Schafer, Konstan, & Riedl, 1999; Häubl & Trifts, 2000; Schafer, Konstan, & Riedl, 2001; Burke, 2002; Konstan & Riedl, 2012).

Since its advent in 1995, Amazon.com has deployed a wide variety of recommender features, ranging from such algorithmic approaches as collaborative filtering to non-algorithmic, non-personalized approaches as consumer reviews (Schafer, Konstan, & Riedl, 1999; Schafer, Konstan, & Riedl, 2001). One side of the equation is the drive for mass customization through the use of recommender systems, i.e. providing each customer with a unique experience that is tailored for them (Linden, Smith, & York, 2003; Schafer, Konstan, & Riedl, 2001). In effect, Amazon.com store changes thoroughly based on the current user interests (Linden, Smith, & York, 2003). Based on such measures as click-through and conversion rates, personalized recommendations have been shown to be a much more effective way to convert browsers to customers than non-personalized, generic product placements, top-seller lists, or hand-generated cross-sell lists (Linden, Smith, & York, 2003; Konstan & Riedl, 2012). Personalization is also seen as a means to build long-term relationship with the customer (Schafer, Konstan, & Riedl, 2001).

The other side of the equation is eWOM, or community opinion and critique (Schafer, Konstan, & Riedl, 2001), that has become a crucial source of information for consumers (Racherla & Friske, 2012). eWOM was part of Amazon.com's arsenal from early on, and they have taken it to such a level that today the company even earns additional revenue from selling customer reviews to other companies (Mudambi & Schuff, 2010).

Amazon.com has also quickly integrated and experimented with many concepts that research has suggested as significant; e.g. explanations and possibilities to adjust the recommendations generated (Schafer et al., 2007).

All in all, recommender systems are today integral to e-commerce and still continue to shape our online consumption experience (Schafer, Konstan, & Riedl, 1999; Said et al., 2012). The significance of recommenders to e-commerce is only likely to grow.

## 9.2 NEWS RECOMMENDING

The advent of the Internet has changed and continues to change news reading habits (Liu, Dolan, & Pedersen, 2010; Cleger-Tamayo, Fernández-

Luna, & Huete, 2012). The number of news items available from a plethora of sources to readers even without subscriptions has exploded and is overwhelming news consumers (Das et al., 2007; Liu, Dolan, & Pedersen, 2010; Cleger-Tamayo, Fernández-Luna, & Huete, 2012). The challenge is to help readers to find news items that interest them, and, consequently, news recommending has become an active area of research (Das et al., 2007; Liu, Dolan, & Pedersen, 2010; Cleger-Tamayo, Fernández-Luna, & Huete, 2012). The key again is serendipity, as in many cases news consumers do not know what news is out there and, at the same time, news quickly becomes, well, yesterday's news (Das et al., 2007; Liu, Dolan, & Pedersen, 2010). Often, in fact, news sites are browsed for interesting items (rather than for some specific news item) with the attitude of *"Show me something interesting"* (Das et al., 2007).

Online news recommending can be seen as taking place in three ways, 1) within a news provider's own service, e.g. *The New York Times*[41] website has a recommender system that suggests salient items to readers who have logged in (and TopN recommendations to those who have not) and the *Washington Post* is also experimenting with personalized news (Krakovsky, 2011); 2) through news aggregator sites, e.g. Google News[42], Yahoo! News[43], Digg[44], and Ampparit[45] (a popular Finnish news aggregator site), where the sites aggregate and recommend news that are provided by separate news sites; and 3) person-to-person and person-to-many recommending that takes place e.g. through email and social networking sites. The first two cases have been discussed more in literature and these topics are briefly reviewed here, while the third case, discussed in Publication III, has not received as much research interest.

When it comes to news provider services or aggregator services, recommendations can be delivered in different ways, e.g. on the website as users browse or by RSS feed (Krakovsky, 2011; Cleger-Tamayo, Fernández-Luna, & Huete, 2012). Besides lists of recommended news items, recommenders are also one way to personalize the front page and overall experience (Cleger-Tamayo, Fernández-Luna, & Huete, 2012). The goal is, in a sense, to create a personal newspaper for each visitor (Liu, Dolan, & Pedersen, 2010; Krakovsky, 2011).

Given the dire financial state that many traditional news providers, especially newspapers, find themselves facing today, there is also a financial question of survival involved in these efforts (Krakovsky, 2011; Granger, 2013). Giving content for free online has not helped the bottom-line of news makers (Granger, 2013)—in hindsight, predictably enough.

41 http://www.nytimes.com/
42 http://news.google.com/
43 http://news.yahoo.com/
44 http://www.digg.com/
45 http://www.ampparit.com/

Personalization through recommenders is seen to offer great business potential (Krakovsky, 2011). Tailoring the experience to provide a better user experience is seen as increasing subscriptions because, in the words of Frons, the chief technology officer at *The New York Times*, "…people who read the product more frequently and more thoroughly are more likely to subscribe to it…", and because it makes, in the words of Benton, the director of Harvard University's Nieman Journalism Lab, "the page … a more valuable piece of property to an advertiser" (Krakovsky, 2011).

While whether or not personalization turns out to be the metaphorical silver bullet to save the journalism business is very much an open question (Krakovsky, 2011), the recent, $250 million acquisition of the money-hemorrhaging *Washington Post* by Amazon's CEO Jeff Bezos on October 1st, 2013, has the world guessing how he plans to turn the old, venerable ship around (Farhi, 2013; O'Toole, 2013). Bezos has suggested that what has worked for Amazon could also work for the newspaper (Solsman, 2013) but has also stated that, "There is no map, and charting a path ahead will not be easy" (O'Toole, 2013). Then again, perhaps the photograph accompanying the CNET news story (Solsman, 2013) can provide a hint; in it, Bezos is introducing Kindle in 2012 and behind him is the slogan: "We want to make money when people <u>use</u> our devices, not when they <u>buy</u> our devices."

On the other hand, there are already some success stories; with recommenders, Yahoo! has increased its click-through rate on articles by more than 160% in comparison to what human editorial judgment reached (Krakovsky, 2011). At the same time, big players are cautiously moving forward in order to not compromise user trust; as Ramakrishnan, Yahoo!'s chief scientist for search and cloud computing, says, "The footprint of what we do is very broad"; e.g. in 2011, more than 100 million users visited Yahoo!'s home page (Krakovsky, 2011).

### News Domain Characteristics

The extreme velocity of online news makes it challenging to collect enough data to make recommendations (Krakovsky, 2011). News items need to be recommended as soon as they are published—immediacy and freshness are essential—and collecting user preference data to use in e.g. collaborative filtering is not practical, as such systems cannot recommend news that have not been read by other users and collecting data is, in any case, too slow a process (Liu, Dolan, & Pedersen, 2010; Krakovsky, 2011; Cleger-Tamayo, Fernández-Luna, & Huete, 2012).

The problem of slowness is compounded by the fact that few users in the domain of news are willing to contribute explicit evaluations of news articles (Liu, Dolan, & Pedersen, 2010; Cleger-Tamayo, Fernández-Luna, & Huete, 2012). At the same time, accurate user profiles are necessary to make personalized recommendations (Liu, Dolan, & Pedersen, 2010). As with ratings, few users are willing to exert effort to build profiles in other ways (Liu, Dolan, & Pedersen, 2010). As a result, implicit measures need to be used; e.g. in most cases, opening of a news article by clicking is considered a positive vote for it (Das et al., 2007; Liu, Dolan, & Pedersen, 2010; Cleger-Tamayo, Fernández-Luna, & Huete, 2012). However, this approach gives little information about negative interests while explicit preference data, e.g. a five-star scale, also provides information about dislikes (Das et al., 2007).

Such an approach naturally results in noisy data (Das et al., 2007). However, as far as Google News is concerned, Das et al. (2007) assume that the snippets provided with links keep the data relatively clean, as users can easily pre-judge which links to click—in fact, the worry is that people get everything they want from the snippet and do not ever click the link even though they found the news item interesting.

Using social approach to eliciting explicit data has great potential, as the success of Digg.com underlines. Clearly, at least under certain conditions, users are willing to provide preference data in news aggregation services. The effectiveness of the collective decision making on Digg is evident in the fact that when Rumsfeld resigned in 2006, the news item was on Digg's front page within three minutes while it took twenty minutes longer before it appeared on Google News. (Lerman, 2007)

A further complication in news recommending is that user interest changes with events (Liu, Dolan, & Pedersen, 2010; Cleger-Tamayo, Fernández-Luna, & Huete, 2012). There are two types of interests when it comes to news, short-term and long-term interest (Liu, Dolan, & Pedersen, 2010). *Short-term interest* changes quickly and is related to currently hot topics, while *long-term interest* reflect user interests in the similar way as in other domains (Liu, Dolan, & Pedersen, 2010).

Analyzing log data from Google News, Liu, Dolan, and Pedersen (2010) found that the user interest does change over time and that there are news trends. General click distribution reflects news trends, corresponding typically to important news events, and these trends tend to be local in the sense that different locations have different news trends. The user's news interest corresponds to a certain extent to the news trend in their location. Consequently, news recommendations need to be tied both to the user's long-term interests and also to the local news trends so that the user does not miss news items that interest them, even though they do not strictly match their long-term interests. (Liu, Dolan, & Pedersen, 2010)

Finally, there are strict response time requirements—e.g. on Google News, the recommendation engine has only "a few hundred milliseconds" to generate recommendations (Das et al., 2007). At the same time, there might be terabytes of daily log data, and, unlike in e.g. e-commerce where some calculating can be done offline as old data can be used (Linden, Smith, & York, 2003), in news recommending, calculating must be undertaken nearly in real-time (Krakovsky, 2011). This does not mean, of course, that prior knowledge is not used but, with new items rolling in and freshness being an important criterion, challenges are also significant at the performance level when it comes to news recommenders (Krakovsky, 2011).

### *The New York Times* Approach and Maintaining Editorial Voice

*The New York Times* bases its personalization on a system of semantic tags that are added manually. These semantic tags allow the recommendation engine to know what each news item is about and the system keeps track of the visitor reading habits based on these tags. *The Times* does not employ collaborative filtering and has, therefore, no way to learn the latent interests of its users. Also, because the system bases its recommendations on articles, it cannot recommend stories on topics that *The Times* has not previously covered. (Krakovsky, 2011)

At the same time, there is a method to the madness. At the core of the system's aims is to provide the right kind of user experience that the users expect and appreciate. A newspaper of *The Times*' caliber is not appreciated simply for individual articles but for its news judgment and its editorial voice. At the time of automatization, newspapers have struggled to keep their traditional role of being arbiters of news, and, while recommenders aim at providing people what they want, the role of a newsperson is also to give the reader the news they did not know they need. (Krakovsky, 2011)

In effect, if the reader is only given the news items they want, there is a risk of overfitting and resulting balkanization (Krakovsky, 2011; Schwind, Buder, & Hesse, 2011; Konstan & Riedl, 2012). People may end up only reading news that support their views and not be exposed to opposing views that might give them food for thought (Schwind, Buder, & Hesse, 2011; Konstan & Riedl, 2012). The consequences for democracy and society at large might not be benign (Krakovsky, 2011; Schwind, Buder, & Hesse, 2011). Traditionally, journalists have seen themselves as playing a crucial role in democratic society, and rampant automatization can be seen as a threat to this (Krakovsky, 2011).

### Google News Approach: Hybrid Recommender

Unlike *The New York Times*, Google News, a leading global news aggregation website, does not worry about editorial voice issues and aims to perfect its algorithmic approach instead. Users who log in and have enabled history can get personalized news recommendations. (Liu, Dolan, & Pedersen, 2010)

As discussed above, the collaborative approach, i.e. using the opinions of like-minded neighbors to generate recommendations without considering the content of the item, does not work due to the news recommending domain characteristics. Content-based systems, on the other hand, generate recommendations based on item content and in news recommending can be divided into those systems that take the user profile into consideration and those that base their recommendations on the current item of the user (Cleger-Tamayo, Fernández-Luna, & Huete, 2012). If collaborative filtering is employed in this context, it is typically combined with a content-based approach to deal with cold-start and first-rater problems (Liu, Dolan, & Pedersen, 2010; Cleger-Tamayo, Fernández-Luna, & Huete, 2012).

Previous to 2010, the Google News recommendation system used a collaborative filtering approach (Das et al., 2007; Liu, Dolan, & Pedersen, 2010). This led to a wait of several hours before enough user preference data could be implicitly collected, meaning that breaking news could not be recommended (Liu, Dolan, & Pedersen, 2010). Moreover, the collaborative filtering approach did not account for individual differences sufficiently; e.g. entertainment news was also recommended to users who never viewed them because entertainment stories in general are very popular, and so non-interested visitors ended up having a lot of neighbors who were interested in entertainment news (Liu, Dolan, & Pedersen, 2010). In effect, collaborative filtering works better in non-urgent tasks, e.g. recommending news articles from an archive (Cleger-Tamayo, Fernández-Luna, & Huete, 2012).

Because there are both long-term and short-term news interests, Liu, Dolan, and Pedersen (2010) took user profiles as a starting point for their approach to the Google News recommender system. With profiles, users with no interest in entertainment news would not get such recommendations even if their neighbors were interested in them. The profiles were based on user activity rather than explicit preference data (for the reasons discussed above). The approach uses a hybrid recommender that combines collaborative filtering with a content-based approach that takes into consideration user profiles to create a personal newspaper for each reader. Moreover, Liu, Dolan, and Pedersen (2010) developed a Bayesian model to combine trends and user interests to make sure that news that are interesting to the user, but do not match their long-term interests, still get recommended to them.

Liu, Dolan, and Pedersen (2010) tested the new approach on a subset of the live traffic of Google News. The results showed a significant improvement over the previous collaborative filtering approach. However, interestingly, while the test group clicked significantly more recommended items, the total number of clicked items did not change. It seems that the total number of news items a user is willing to view is constant; the improved recommender simply attracted users away from non-personalized sections. In effect, users in the test group spent less effort in finding interesting news. Also, test group members visited the Google News site more often, increasing overall traffic to the site.

## 9.3 RECOMMENDER SYSTEMS IN e-LEARNING

Recommender systems are seen as offering a great potential in e-learning, and there is an increasing interest in them in that domain (e.g. Tang & McCalla, 2004a; Hage & Aïmeur, 2008; Drachsler, Hummel, & Koper, 2009; Ghauth & Abdullah, 2010; Manouselis et al., 2011; Buder & Schwind, 2012). As in other domains, recommender systems are seen in e-learning as offering at least a partial solution to information overload issues as decision support systems (Farzan & Brusilovsky, 2006; Tang & McCalla, 2009). In addition, using recommender systems in e-learning may actually aid the learning process itself, as learning today is considered an active and constructive process and recommenders may lead to participation, collaboration, and therefore deeper involvement and elaboration (Recker, Walker, & Lawless, 2003; Stefani, Vassiliadis, & Xenos, 2006; Buder & Schwind, 2012). However, the development efforts are slowed down by the fact that while many approaches to using recommender systems are being considered, we have few experiences of recommender systems having been evaluated with trials involving actual use contexts and genuine users (Drachsler, Hummel, & Koper, 2009; Santos & Boticario, 2010; Manouselis et al., 2011).

Simultaneously, there is a common misconception that recommender systems in e-commerce and e-learning are governed by the same principles (Stefani, Vassiliadis, & Xenos, 2006). In fact, in spite of similar purposes, moving recommenders from one domain to another is challenging, as recommender systems are strongly domain dependent and recommendation tasks in different domains may differ or have different emphases, in addition to differing contextual particularities (Tang & McCalla, 2004a; Drachsler, Hummel, & Koper, 2009; Ghauth & Abdullah, 2010; Santos & Boticario, 2010; Buder & Schwind, 2012). For example, in e-learning, the objectives of the recommender system designers and their users are the same, i.e. learning[46], while in e-commerce the objectives of the two are not; in e-commerce, designers aim at maximizing selling while users are not looking to maximize buying (Stefani, Vassiliadis, & Xenos, 2006; Buder & Schwind, 2012).

[46] …although it may be argued that the student typically aims to accomplish the required learning with minimal effort, while the teacher may wish that the student would read more widely on the topic.

There is a contextual richness in the educational domain that needs to be taken into account and that provides many opportunities for employing recommenders in e-learning (Stefani, Vassiliadis, & Xenos, 2006; Santos & Boticario, 2010). While the user tasks and goals identified by Herlocker et al. (2004), e.g. *annotations in context* and *recommending in sequence*, are, by and large, valid for e-learning recommenders, they serve only as a starting point (Drachsler, Hummel, & Koper, 2009; Manouselis et al., 2011).

Learning contexts to which recommenders are being introduced can vary widely; what is recommended varies (e.g. courses vs. study materials) as does the function of the recommender (e.g. supporting tutor vs. replacing tutor) and the type of learning (formal vs. informal) (Farzan & Brusilovsky, 2006; Drachsler, Hummel, & Koper, 2009; Ghauth & Abdullah, 2010; Granić & Adams, 2011). In e-learning, not only can objects—e.g. learning materials, web resources, or courses—be recommended but recommendations can also be actions aimed at achieving an educational goal, e.g. doing an exercise to strengthen a weak area, reading a post, or posting a reflecting message on a topic, and engendering interactivity (Santos & Boticario, 2010).

Recommenders have been studied in a broad range of educational contexts for numerous different purposes (Buder & Schwind, 2012), ranging from recommending resources on the web, lecture notes, assignments, and other learning objects to recommending entire courses and foreign language lessons in such contexts as libraries, workplace learning, and mobile learning (Farzan & Brusilovsky, 2006; Tang & McCalla, 2009; Buder & Schwind, 2012).

Although not widely discussed in e-learning literature, the fact that recommender systems are social by nature has important implications to e-learning (Buder & Schwind, 2012). Recommenders can provide social presence, a sense of not being alone in the space, and build a sense of community among frequent visitors (Chalmers et al., 2004; Svensson, Höök, & Cöster, 2005; Kumar & Benbasat, 2006; Mudambi & Schuff, 2010); in fact, "the mere existence of reviews established social presence, and that online, open-ended peer comments can emulate the subjective and social norms of offline interpersonal interaction" (Mudambi & Schuff, 2010). Social presence can turn activity into interactivity (Leino, 2013), and as active construction of knowledge is the current trend in education, recommenders offer possibilities of turning activity and interactivity into learning episodes (Recker, Walker, & Lawless, 2003; Stefani, Vassiliadis, & Xenos, 2006; Buder & Schwind, 2012).

While the goal for recommenders in e-commerce is to facilitate purchasing, helping users find suitable products and choose which one to purchase, recommenders in e-learning are aimed at facilitating learning and competence development, as well as supporting learning activities and processes (Stefani, Vassiliadis, & Xenos, 2006; Drachsler, Hummel, & Koper, 2009; Manouselis et al., 2011; Buder & Schwind, 2012). Consequently, the recommendations in e-learning should be guided by educational objectives rather than by user interests alone (Drachsler, Hummel, & Koper, 2009; Tang & McCalla, 2009; Granić & Adams, 2011; Santos & Boticario, 2010). Therefore, although Burke (2002) argues that it is interestingness and usefulness (in addition to individualization) that separates recommender systems from retrieval systems, this does not actually entirely apply in the domain of e-learning where interestingness loses part of its importance (Tang & McCalla, 2004a, 2009; Santos & Boticario, 2010); e.g. learners are ready to read papers recommended to them that are not interesting but are pedagogically grounded—the user goal is, after all, learning (Tang & McCalla, 2004a, 2009).

Depending on the context, recommenders in e-learning also need to provide longer-term support to learners than recommenders in e-commerce where the point is typically to provide solid commercial arguments that allows the purchasing decision to be made (Drachsler, Hummel, & Koper, 2009; Manouselis et al., 2011; Buder & Schwind, 2012). Consequently, recommender systems in e-learning may need to consider at what point to recommend which item to the particular learner, i.e. the order in which to recommend items, the learner's background knowledge, and the suitability of the material to the learner (e.g. not recommending highly technical papers to beginners even if the material itself is good) (Tang & McCalla, 2004a; Santos & Boticario, 2008; Drachsler, Hummel, & Koper, 2009; Buder & Schwind, 2012). In fact, instead of recommending individual items, there may be a need to recommend *learning paths* that also consider "sequential dependencies among learning items" (Buder & Schwind, 2012; see also Stefani, Vassiliadis, & Xenos, 2006; Drachsler, Hummel, & Koper, 2009). Pedagogically, the recommended items should be in Vygotsky's *zone of proximal learning*, i.e. slightly above the learner's current level of competence to allow scaffolding (Drachsler, Hummel, & Koper, 2009; Manouselis et al., 2011; Buder & Schwind, 2012). The challenge is compounded by the fact that not only are such aspects as the learner's background knowledge and material suitability difficult to assess, they also change over time as learning takes place (Drachsler, Hummel, & Koper, 2009; Buder & Schwind, 2012). In effect, e-learning recommenders need a lot of context-awareness (Buder & Schwind, 2012). As a result, Drachsler, Hummel, and Koper (2009) consider the recommendation goal to be more complex in the e-learning domain than in the e-commerce domain.

Moreover, it is important to recognize that learning itself is not a uniform domain; for example, it can be divided into formal and informal learning (Drachsler, Hummel, & Koper, 2009). *Formal* learning refers to learning in such institutes as universities and schools that leads to accreditation and provides domain expertise that aims to guarantee quality. In contrast, *informal* learning is less structured and does not lead to accreditation. Also, it may be intentional but is often unintentional, taking place as part of everyday activities at home and work. Consequently, in supporting learning with recommender systems in formal e-learning, learning plans and existing structures can inform the implementation, while in informal learning, information comes from many sources, many of them non-institutional, and there is "an absence of maintenance of metadata and of predefined semantic relationships between" learning activities and objects. (Drachsler, Hummel, & Koper, 2009)

Consequently, Drachsler, Hummel, and Koper (2009) see the main recommendation goals in informal learning as *structuring learning activities in a pedagogical way* to support competence development and *suggesting emerging learning paths*, i.e. recommending sequences of learning activities that lead to reaching a specific learning goal. Suggesting learning paths could allow learners to choose between the most efficient (save time) and most effective (maximize quality) paths to help them choose the appropriate goal for themselves (Drachsler, Hummel, & Koper, 2009).

Connected to the concepts of formal and informal learning are the two perspectives, that of top-down and that of bottom-up. *Top-down* perspective uses pre-defined educational meta-data and filtering decisions coming mainly from educators to inform recommendations, while *bottom-up* perspective approaches recommending from the perspective of learner-produced information, e.g. tags, ratings, and behavioral data (Drachsler, Hummel, & Koper, 2009; Santos & Boticario, 2010). Naturally, the two can also be combined when both types of data are available, but informal learning often lacks the top-down data, making it dependent of learner-generated data.

Moreover, it is not only learners that are in need of recommendations; educators also need support for various tasks as they face a chronic shortage of time, especially as using online resources is increasingly emphasized (Recker, Walker, & Lawless, 2003; Manouselis et al., 2011). Educators need support to quickly and easily find high-quality instructional resources for preparing lessons in addition to needing support in delivering the lessons (i.e. teaching), as well as in the evaluation and assessment phase (Recker, Walker, & Lawless, 2003; Manouselis et al., 2011). These tasks increase the variety of recommendation goals in e-learning (Manouselis et al., 2011).

In effect, the e-learning domain in and of itself requires many different approaches to recommender systems; just like one approach from one domain cannot be directly used in another domain, a successful recommender system in one kind of educational context cannot directly be employed in another educational context (Ghauth & Abdullah, 2010; Leino, 2012a).

Given all the complexity and changing user model as a result of learning (Drachsler, Hummel, & Koper, 2009), it is perhaps no wonder that adaptive systems have enjoyed scant success in e-learning (Stefani, Vassiliadis, & Xenos, 2006; Granić & Adams, 2011). One of the important reasons is the difficulties related to user modeling (Drachsler, Hummel, & Koper, 2009; Granić & Adams, 2011). In e-learning, it is challenging to collect enough information to initiate the user model (Granić & Adams, 2011), and, consequently, researchers have been forced to consider e.g. such solutions as generating artificial learners to provide recommendations (Tang & McCalla, 2004b).

### Recommender Systems and Learning

Buder and Schwind (2012) consider such aspects of recommender systems as collective responsibility, collective intelligence, user control, guidance, and personalization to make a good fit with learning. Recommender systems shift the focus away from experts to the peer collective in the same way as there is currently a trend to move away from teacher-centered learning towards learner-centered learning. In both fields, the trend is towards flat hierarchies and there is an assumption that peer efforts lead to better results. (Buder & Schwind, 2012)

Recommender systems rely on collective intelligence, and so the generated content is not meaningfully traceable back to any individual but to the collective behavior of the collective (Buder & Schwind, 2012). It has also been argued that recommender systems exhibit collective intelligence (Malone, Laubacher, & Dellarocas, 2010; Buder & Schwind, 2012). In the same way, in educational science, there is a notion of group cognition according to which the artifacts and discussions of a learner group cannot be meaningfully traced back to any individual member of the group (Buder & Schwind, 2012). In effect, learning groups and recommender systems can be seen as exhibiting emergent properties (Buder & Schwind, 2012).

The recommendations that recommender systems generate are suggestions and in that sense different from e.g. a mandatory book on a course; the user has control over whether or not to follow recommendations. In effect, recommenders preserve user autonomy, and therefore are similar to modern constructivist epistemology that underlines the importance of self-regulation, exploration, and autonomy in learning. In other words, recommenders provide guidance that learners may or may not take into consideration. Guidance is also central in learning where too much autonomy has proved to be problematic and some kind of structuring, be it explicit or implicit, is seen as having a positive effect. The question is to find the balance between autonomy and guidance. (Buder & Schwind, 2012)

Finally, recommender systems can provide personalization, adapting the suggestions to the individual learner (Buder & Schwind, 2012). Likewise, in educational science, there is a consensus that uniform instruction does not serve all learners equally and therefore instructional material should be adapted to the level of knowledge and the abilities of the individual learner (Stefani, Vassiliadis, & Xenos, 2006; Buder & Schwind, 2012).

In effect, Buder and Schwind (2012) consider recommender systems to theoretically have, at least, a great "potential to leverage learning processes," as the educational principles are embedded in their very nature.

### Learner Input Elicitation as a Learning Opportunity

Recommender systems use explicit or implicit (or both) feedback from the user community in order to distill collective wisdom (Farzan & Brusilovsky, 2006). While explicit feedback is often considered more reliable, getting learners to provide it has proven challenging also in the e-learning domain (Recker, Walker, & Lawless, 2003; Farzan & Brusilovsky, 2006). In e-commerce, the goal is often not to distract the user from their main task and to keep the burden of preference elicitation as small as possible, leading to an emphasis on implicit collection of preference data (Buder & Schwind, 2012; Konstan & Riedl, 2012). In fact, Pu, Chen, and Hu (2012) give the easiness of the preference elicitation process as a guideline to designing recommenders. Also, it has been suggested that implicit preference data might be more objective as it is not biased by users trying to behave in a socially desirable way (Buder & Schwind, 2012).

However, there are strong arguments for using the explicit collection of preferences in e-learning. Constructivist learning theories, as discussed, emphasize the building of knowledge through social interaction and collaboration, as learning is seen as an active, constructive process, and the explicit collection of preference data is a way to involve and engage learners, making them deliberately reflect and elaborate on issues and objects, allowing them to thereby gain information literacy skills (Recker, Walker, & Lawless, 2003; Stefani, Vassiliadis, & Xenos, 2006; Buder & Schwind, 2012). In effect, providing explicit feedback can be seen as an opportunity to learn and therefore as an aid to the learning process rather than as a burden (Recker, Walker, & Lawless, 2003; Buder & Schwind, 2012).

Also, the explicit collecting of user preferences is associated with user control and user autonomy that is seen both as leading to user satisfaction and trust and as associated with modern learning approaches (McNee et al., 2003; Buder & Schwind, 2012). Moreover, explicit feedback collecting helps make the contribution to the community visible, fueling collaborative spirit (Buder & Schwind, 2012). Although there seems to be a need to provide incentives for learners to contribute explicit feedback, course credit can be used for this (Recker, Walker, & Lawless, 2003).

#### Confirmation Bias: Need to Recommend What Is Not Wanted

Human beings are prone to *confirmation bias*, i.e. the tendency to seek information that confirms their views and preferences, and to discard information that does not support them (Buder & Schwind, 2012). As discussed in the context of news recommending, recommenders can make this even worse by recommending only items that users are likely to like, potentially leading to balkanization where we only get to see viewpoints that are similar to ours (Schwind, Buder, & Hesse, 2011; Buder & Schwind, 2012; Konstan & Riedl, 2012). In education, information diversity is necessary for open-minded and critical thinking (Buder & Schwind, 2012). Consequently, in e-learning and potentially also in other domains, there is a need to also recommend items that provide viewpoints that challenge our preconceptions (i.e. the recommender has to understand our preferred views and confront them with evidence to the opposite) and the viewpoints that the items we have been using promote (e.g. if a learner is using a particular theory in an essay, it would be good to provide them with papers that are critical of that theory) (Buder & Schwind, 2012). Schwind, Buder, and Hesse (2011) showed that confirmation bias can, in fact, be reduced if preference-inconsistent information is highlighted; however, simply making it available was not enough.

Again, it becomes clear that using user interests and tastes as the sole guide to recommending items is not a suitable approach (Drachsler, Hummel, & Koper, 2009; Tang & McCalla, 2009; Santos & Boticario, 2010; Buder & Schwind, 2012). In fact, the most useful recommendations may occasionally actually be for items that learners do not like (Buder & Schwind, 2012).

#### Evaluating Educational Recommenders

Unlike in the case of many other domains, in educational recommenders there are no standardized data sets for evaluating algorithms or even standardized evaluation procedures, for that matter (Drachsler, Hummel, & Koper, 2009). Likewise, when it comes to commercial recommenders, such actions as purchasing can be used in evaluating recommenders; in contrast, in e-learning there are not always clear-cut, instant learning events that could be used to measure the effectiveness of the recommender (Buder & Schwind, 2012). Consumption of a learning object, for example, does not tell if the level of the object was suitable to the learner (Buder & Schwind, 2012). Satisfaction in educational recommenders is connected to learning, to mastering competencies, and therefore needs to be considered in terms of how well the system supports the learning process (Drachsler, Hummel, & Koper, 2009).

Consequently, Drachsler, Hummel, and Koper (2009) suggest that we need to "mix technical evaluation criteria with educational research measures"; there is a need to evaluate if learners benefit from using the system—whether the system "makes learning more effective, efficient, or more attractive." They suggest the following framework for evaluating educational recommender systems (Table 2):

| ***Measurement*** | ***Parameters*** |
|---|---|
| **Technical measures** | Accuracy |
| | Coverage |
| | Performance |
| **Educational measure** | Effectiveness |
| | Efficiency |
| | Satisfaction |
| | Drop-out rate |
| **Social network measures** | Variety |
| | Centrality |
| | Closeness |
| | Cohesion |

**Table 2**. Evaluation framework for educational recommender systems (Drachsler, Hummel, & Koper, 2009).

In educational research, *effectiveness* ("total amount of completed, visited or studied learning activities"), *efficiency* (time needed to reach the learning goal), *satisfaction* ("the individual satisfaction of the learners with the given recommendations"), and *drop-out rate* are commonly used measures. Social network measures aim to measure the benefit that the contributions of the learners provide the learner network as a whole, i.e. they are measures related mainly to informal learning networks. *Variety* measures the level of emergence of the most successful learning paths. *Centrality* measures the connectivity of a learner in the network. *Closeness* measures how close a learner is to all the other learners in the network, representing "the ability to access information direct or indirect through the connection to other network members." Finally, *cohesion* measures how strongly "learners are directly connected to each other by cohesive bonds." (Drachsler, Hummel, & Koper, 2009)

While the framework suggested by Drachsler, Hummel, and Koper (2009) has by no means become a standard to evaluating educational recommenders, it is introduced here as we are not aware of any other framework for this particular domain.

# 10 Discussion

The publications cover three domains where recommender systems have been employed, namely, e-commerce (papers I and II), news recommending (paper III), and e-learning (papers IV–VII). We first look at the general contribution that the papers make and then look at the contributions in each domain separately.

## 10.1 General Contribution: Insight into Actuality of Use

Although recommender systems have been a subject of great research interest, user-centric aspects, such as user experience, have been studied to a much lesser degree than system-centric aspects (Herlocker et al., 2004; Cremonesi et al., 2011; Knijnenburg, Willemsen, & Kobsa, 2011; Buder & Schwind, 2012; Knijnenburg et al., 2012; Pu, Chen, & Hu, 2012). Given that researching user experience and similar constructs requires developing and running or having access to a running system and requires a user-community that uses the system, the fact that limited empirical work has been done from a user-centric perspective is understandable (Herlocker et al., 2004; Svensson, Höök, & Cöster, 2005; Redpath et al., 2010; Cremonesi, Garzotto, & Turrin, 2012; Konstan & Riedl, 2012). The fact that user-centric issues are more complex and difficult to articulate and operationalize has further contributed to their neglect (Cremonesi et al., 2011).

However, although such neglect may be understandable and valuable contributions have been made through system-centric research, this one-sided emphasis has resulted in us having relatively little understanding about how users perceive recommenders and recommendations and how users are effected by them (Cremonesi et al., 2011; Cremonesi, Garzotto, & Turrin, 2012a; Buder & Schwind, 2012; Racherla & Friske, 2012). Moreover, recommenders are becoming increasingly common and important, and therefore "the ability to characterize user experience and users' affective attitudes toward the technology has become extremely important" (Pu, Chen, & Hu, 2011).

In addition, many user-centric studies that have been conducted have used students acting as users doing tasks that are not engaging or involving, i.e. using non-authentic contexts (Grabner-Kräutera & Kaluscha, 2003; Sillence & Briggs, 2007). Authenticity is important also because researcher-defined tasks may preclude the use of certain types of behaviors or strategies and because users appear to employ different strategies when they carry out tasks that they have defined themselves than tasks that the researcher has defined (Thatcher, 2006).

Moreover, studies that involve large numbers of participants tend to be questionnaire studies, meaning that say-do problems cannot be ruled out (Kobsa, 2007). For example, the oft-quoted study by Ozok, Fan, and Norcio (2010)—for example, Pu, Chen, and Hu (2011) recommend using its model next to their own framework—had 131 participants, "mostly full-time or part-time college students from a university", fill out an online questionnaire and Pu, Chen, and Hu (2011) evaluated their user-centric evaluation framework using an online questionnaire that was filled out by 239 participants (of whom 202 were female and 150 students). Incidentally, both studies used Likert-style scales with labels for all options and treated them as interval scales for calculating statistical values, a somewhat questionable practice (Allen & Seaman, 2007).

Of the publications here, only Publication III is based purely on questionnaire data, with most participants being students, as the study was meant for setting hypotheses for further studies. In all other publications, participants were genuine users in a genuine use context and used genuine systems.

Moreover, both observation, i.e. behavioral data based on either direct observation while the use took place (papers I and II) or use log data (papers IV, V, VI, and VII), and subjective responses, collected by interviews (papers I, II, and VII) or questionnaires (papers I, II, VI, V, VI, and VII), were collected and used in analysis, as suggested by Herlocker et al. (2004) and, more recently, Knijnenburg et al. (2012). This triangulation of collected data types and data collecting methods[47] lends credence to the

[47] There are typically considered to be four forms of triangulation, 1) data triangulation (collecting different types of data and collecting data with more than one sampling strategy), 2) investigator triangulation (the use of more than one researcher in collecting and interpreting data), 3) theoretical triangulation (the use of more than one theoretical view in interpreting data), and 4) methodological triangulation (the use of more than one method to collect data). Combining qualitative and quantitative approaches is also often considered to be a form of triangulation. (Bryman, 2004)

results, as using only one method of collecting data or collecting only one type of data can expose the findings to any limitations associated with the method or the used application of it (Bryman, 2004). In addition, while observation data tells *what* happened, it provides little information concerning *why* it happened—what were the reasons for users to adopt certain behaviors or strategies (Thatcher, 2006). While subjective response data, or explicit data, provides valuable insights into the reasons and motivations of users, collecting it is not straightforward (see Sections 7.3 and 7.5 for some reasons why collecting explicit response data is challenging). Moreover, even if we had ideal research instruments that do not bias responses and best-intentioned participants who wish to convey us the truth as they know it, we know that what people say and what people actually do are not necessarily the same thing[48]. To reduce the effect of the say-do problem, we balanced saying (subjective responses) with doing (observation data) in all but one publication (Table 3). Similarly, we collected and used both qualitative and quantitative data in all studies[49]. As a result, we are confident that these publications here afford us genuine insights into user experiences.

| Publication | Area | Observation | Subjective response |
|---|---|---|---|
| I, II | E-commerce | Over-the-shoulder and video | Interview and questionnaire |
| III | News | - | Questionnaire |
| IV, V, VII | E-learning | Use log data | Questionnaire |
| VI | E-learning | Use log data | Questionnaire and interview |

**Table 3**. The publications by the types of data used.

Much research on recommender systems has focused on certain aspects of the systems, isolating them from the whole of the system, or focusing on recommender systems in isolation from the ecosystem to which they have been embedded. The user, however, typically faces recommenders as part(s) of complex wholes, not as isolated aspects or systems, and consequently the complex whole is what the user has to cope with and what gives birth to the user experience (Juvina & van Oostendorp, 2004; Freyne et al., 2007). In fact, only recently have researchers in the field of information searching begun to consider "a more holistic view of humans as situated actors within an environment," including digital environments, that "strongly influence their thoughts, processes and behaviors" (Dinet, Chevalier, & Tricot, 2012). One reason for there being so few studies of complex wholes is that studying them is challenging, no matter which methodology is used (Kumar & Benbasat, 2006).

[48] See e.g. Kobsa (2007) for numerous examples of this.

[49] In case of Publication III, both qualitative and quantitative data were collected using a questionnaire. In the case of other publications, much of the qualitative data is based on observation data although some was also collected through interviews or questionnaires.

While studying isolated parts and aspects of ecosystems is necessary, the problem with studying them without looking at the whole is that complex systems result in emergent behavior, as users occupy and transform *spaces* into *places* (see discussion in Section 2.8). Behavior in an online environment is determined to a large part by the interactivity that the setting offers, as humans adapt their behavior and strategies to specific situations and environments (Häubl & Trifts, 2000).

For example, emergent behavior becomes apparent in the recommendations that emerge from the independent decisions and actions of the users in the place (i.e. their understanding of the space); collective decision-making emerges without the collective coming together to make a decision on purpose (Lerman, 2007; Malone, Laubacher, & Dellarocas, 2010; Buder & Schwind, 2012). In effect, complex ecosystems, such as Amazon.com, provide a space for collective intelligence to emerge; individuals form a loosely organized group, e.g. a user community on Amazon.com, doing individually what they do and the result is the aggregate of the individual actions that exhibits intelligence (Malone, Laubacher, & Dellarocas, 2010; Buder & Schwind, 2012). A recommendation on Amazon.com, for example, cannot be traced back to the behavior of any one individual but to the behavior of the user collective (Buder & Schwind, 2012). In effect, the presence and activity of the collective benefits its members (Ling et al., 2005).

However, the only way to study emergent behavior that emerges from using a system is to have actual users use the system (Lerman, 2007)—even if we understand the impact of one aspect of the system, we do not know how that impact is mediated by other aspects without trying it out; e.g. the evaluation framework by Knijnenburg, Willemsen, and Kobsa (2011) places a strong emphasis on understanding mediating effects in user experience.

Again, apart from Publication III, all publications here focus on wholes without isolating recommenders from their natural context and environment. As a result, the publications here make an important contribution to the field by showing how genuine users in a genuine use context and environment use and experience recommender systems. The publications paint a clear picture of user behavior and subjective responses to the environments in their whole complexity, and this alone is a significant contribution to understanding how users see and react to recommender systems in real life, as this has not been studied much by others (Buder & Schwind, 2012).

Publication III used a survey approach to comprehensively study different aspects of news recommending to reveal how news is found online and what kinds of social and other practices have grown around news recommending activity. In effect, its comprehensiveness comes from the scope of the study that also accounted for many aspects that were not directly related to recommender systems, thereby effectively laying bare many social undercurrents of the activity. Consequently, despite surficial differences in comparison to other publications, it also provides an extensive view of the news recommending activity that also considers the context of the activity.

#### Validity in the Studies

At the same time, however, because studying systems in this manner by necessity results in case studies, we have to be careful about generalizing the results too far. The results of these kinds of exploratory studies can be used to build hypotheses for more isolating studies but in and of themselves, they cannot be treated as having great external validity. In effect, the point here is to have validity in the sense of trustworthiness, and using both observation data and subjective response data is the foundation for this. Also, due to using real users in actual use contexts, ecological validity of publications I–II and VI–VII is high.

Also, all the publications here have Finns as participants. When we study phenomena through representatives of one culture, we have to exercise care when generalizing the results to other cultures. For example, culture is, in many ways, a shaping factor in communication; e.g. collectivist cultures, such as South Korea, emphasize relationship-building aspects in communication while individualistic cultures, such as the USA, focus more on information (Goldstein, Martin, & Cialdini, 2007, pp. 196–205). Scandinavian countries tend to fall somewhere in the middle of this axis, amalgamating aspects of both approaches (Kalogeraki, 2009).

## 10.2 E-Commerce: Publications I and II

I. Juha Leino and Kari-Jouko Räihä (2007): Case Amazon: Ratings and Reviews as Part of Recommendations.
   - Strategies for finding items of interest that have emerged
   - Use of recommender features in finding items of interest
   - How items on list pages are selected for closer inspection and what role star ratings play in this
   - Use of reviews in selecting items for purchasing

II. Juha Leino and Kari-Jouko Räihä (2008): User Experiences and Impressions of Recommenders in Complex Information Environments.
   - Role of recommendations in finding items of interest
   - Opportunistic use of recommenders versus strategic use
   - Recommenders and social presence
   - Impact of social presence on behavior

Publications I and II focus on e-commerce and are based on one in-depth study on how users use recommender systems to find and select items and what kinds of use strategies have evolved around recommender systems when they are part of a complex online shopping environment. We used Amazon.com/co.uk as the complex information ecosystem because it provides a rich study ground, being both popular with consumers and widely considered a pioneer and trailblazer in employing various approaches to recommenders in e-commerce (Chalmers et al., 2004; Kumar & Benbasat, 2006; Ozok, Fan, & Norcio, 2010; Pathak et al., 2010).

The study on which the two publications are based had six participants, all male, who had previously purchased at least one book from Amazon (on average, the participants had purchased ten books; the lowest was two books). The study was conducted in 2007. We used applied ethnography, a combination of over-the-shoulder observation with verbal protocol when the participants used Amazon.com/co.uk and conducted a semi-structured interview with each participant after the use. Verbal protocol was used because, while observation allows us to see *what* happens, without verbal protocol we cannot know *why* it happens. After the interview, each participant also filled out a questionnaire.

The study sessions, one per participant, lasted 2–3 hours per person and were videotaped with the camera aimed at the computer screen both to record what was said (verbal protocol and interview) and to allow analyzing how the tasks were done. The sessions took place at the participants' home where they typically used Amazon.

The participants undertook four tasks, the first two of which were used for these papers. Task 1: "*Buy a book (or books) from Amazon. Do not buy a book that you have already decided to buy. Instead, you should find a book that you have not decided to buy beforehand.*" Each participant was given 15 € for purchasing book(s) in Task 1 to make sure that they selected a book they really wanted—no other compensation was given. Therefore, how well they did the task affected what they received from participating; if we had only asked them to find a book to buy, they might easily have used heuristics without deeper involvement in the task[50]. Now participants used on average 23 minutes for finding a book(s) (one participant purchased two books, all others one).

Task 2: "*You have bought a good digital camera and now you would like to buy a photography guide from Amazon. Which one of the books on the list would you buy?*" The list page was pre-constructed and it looked exactly like a list

[50] In comparison, e.g. in Pu, Chen, and Hu's (2011) validation of their framework, the main user task was to "*find an ideal product to buy or experience from an online site employing recommender technology*" in a site of their choice and then fill in the survey; Apple products were raffled among participants.

page on Amazon.co.uk. Pre-constructing was necessary to make sure that the list page included books with high-star rating, low-star rating, no-star rating, and one book with *Search inside* function available. Links led to the actual item pages on Amazon.co.uk.

There are several limitations connected to this study besides the typical limitations that are part of any case study, i.e. the impact of contextual factors. We were limited to only six participants because of our method, and these six participants were all Finnish males. The low number of participants naturally affects the generalizability of the results. In particular, the low number of participants may have resulted in some behavioral strategies that are used in Amazon.com/co.uk not having gotten observed. The observed strategies, therefore, may represent only a subset of all strategies in use. Also, the low number of participants does not allow drawing far-reaching conclusions about the prevalence of the observed strategies. Moreover, men and women are known to exhibit at least some differences as e-commerce customers (Cyr et al., 2007). In effect, while there is a fair amount of ecological validity, the external validity is not necessarily high, and we need to exercise care in generalizing the results.

#### Contributions in Publication I and II

Publication I focused on item-finding and item-choosing strategies of the participants, as well as how reviews and rating affected which item(s) were ultimately chosen. While Publication II partially reviewed the strategy part—we were invited to extend Publication I for this publication—it furthermore discussed the importance of social presence and its impact on the user behavior.

While the importance of the interface is nowadays widely emphasized, prior to our work, effects of interface had received little research interest (Cosley et al., 2003); Cosley et al. (2003) had studied the effect of the rating interface to the ratings given while we studied how the interface affected item-selecting.

Häubl and Trifts (2000) had previously argued that “decision making in complex environments” was characterized by a two-stage process where participants screen a large number of items quickly, selecting some for later in-depth inspection and then do the in-depth comparison of the selected items before the actual purchase decision. However, only two participants in our study used this strategy. We call this *Best item strategy*, as the goal of these two participants was to find the best item (all participants bought non-fiction books). The other four, however, used what we call *Good enough item strategy*, as they scanned the list and, upon spotting a potentially interesting item, moved to its item page, and if it was good enough, they put it into the shopping cart. If it was not good enough, they returned to the list and continued to scan for the next potentially interesting item until they found one that was good enough.

As Häubl and Trifts (2000) had posited, the environment did largely dictate behavior by giving it frames with the interactivity provided; the participants clearly adapted their use to the environment. Amazon offered—and still does—many ways to find books; e.g. searching, recommendations, and categories. Only one participant tried to use categories, and, in practice, all books were found either by recommendations (3) or by searching (4).

In effect, there were clearly two ways to use recommenders in finding books. We call them opportunistic use and strategic use. *Opportunistic use* refers to users stumbling on recommendations while using the system and then using them if something interesting popped up. *Strategic use*, in contrast, refers to intentionally making sure that one gets suitable recommendations; e.g. participants made searchers for items they knew or even already owned with the purpose of using *Customers who…* recommendations to find interesting items on the same topic or went for *Personalized recommendations* directly to see what was recommended to them. However, even today, there is little in terms of support for strategic use approach although it appeared rather common. The participants did not divide into opportunistic users and strategic users, however; many used both approaches, depending on the situation.

Significantly, if we had isolated the recommenders from the larger environment, we would not have been able to observe these two approaches to using recommenders in item-finding. Searching with keywords would not have taken place and its connection to seeding recommenders would not have been noticed. This practice, however, does lend credence to claims that users should be able to easily adjust what is being recommended to them (Pu, Chen, & Hu, 2011).

Customer reviews were crucial to the participants in deciding which book to buy, although they were typically not the sole deciding factor. As literature suggests (e.g. Cui, Lui, & Guo, 2012; Racherla & Friske, 2012), negative reviews had a more pronounced impact although positive ones were also important in the purchasing decision.

Reviews were used in two ways. First, they were seen as a source of information about the books, their content, style, and so on. For this purpose, the star rating of the review was irrelevant to the participants. Second, the participants reflected their needs and expertise level against those of the reviewer to find relevant reviews. In effect, a negative review by an expert in the field did not necessarily matter if the participant was a beginner and vice versa. Also, a glowing review was not important if the needs of the reviewer were very different from the participant's needs.

Only a review by a reviewer who had similar needs, i.e. was similarly positioned vis-à-vis the book, as the participant, the valence of the star rating became meaningful.

In other words, as has been argued in literature afterwards (Hu, Liu, & Zhang, 2008; Castagnos, Jones, & Pu, 2009; Pathak et al., 2010; Pu, Chen, & Hu, 2011), the participants used reviews to reduce uncertainty about the items. Also, literature posits that users prefer guidance from similar people (Chen, 2008; Racherla & Friske, 2012). In fact, in the light of our results, users prefer guidance from people *similarly positioned to the item* as themselves, and this does often indeed mean similarity between the user and reviewer but only as far as relevant aspects go—general similarity was not a relevant factor. For example, a user who is beginning photography would be impressed by a five-star review by another beginner photographer who praises the book as a brilliant guide for a beginner. They would not, however, be discouraged by a one-star review by a professional photographer who considers the book only suitable for an absolute beginner and otherwise absolutely worthless.

The participants carefully read only about 2–3 reviews. The ones that were selected for reading were typically longer ones—participants felt that shorter ones cannot give proper reasoning for their conclusions—and written in matter-of-fact style. There is no clear consensus on whether longer or shorter reviews are more helpful; Mudambi and Schuff (2010) and Chevalier and Mayzlin (2006) see longer reviews as being more helpful as they convey more information and therefore give consumers confidence to act, while Racherla and Friske (2012) see no correlation between the length and the helpfulness of a review, concluding that users prefer reviews "short, sweet and to the point". While our results side more with Mudambi and Schuff (2010) and Chevalier and Mayzlin (2006), this is not to say that the participants would have had patience for rambling reviews. In effect, giving clear reasons for the conclusion does require a certain amount of space but still needs to be clearly written and to the point. We conjecture that a certain length is necessary for giving convincing reasoning and argumentation but this does not translate into the-longer-the-review-the-better-it-is type of heuristic; if we try to correlate length with usefulness, we might get correlation at short and medium length reviews but the longest ones are not necessarily the most helpful ones.

Of course, the question is also partially related to what is used as the definition of *helpful*; typically, the number of *Helpful* votes is used to judge this (Racherla & Friske, 2012) but given the blatant and widespread dishonesty as far as reviews are concerned (Streitfeld, 2012), it is not inconceivable that helpfulness votes are also affected by comparable dishonesty. Also, reviews tend to take extreme positions due to such motivations as *brag and moan* (Hu, Zhang, & Pavlou, 2009), and this may also affect how helpful votes are given. However, helpful votes have not received similar research interest as reviews and other aspects of eWOM.

In any case, our results point out that users do pay attention to the text part of the reviews and do not simply draw conclusions based on the numerical ratings, as literature also indicates (Racherla & Friske, 2012). Also, very evidently, as Racherla and Friske (2012) point out, some reviews have much greater influence on the decision than others.

The participants showed a clear preference for customer reviews over the editorial reviews, as has also been discussed in literature (Chen, 2008; Racherla & Friske, 2012). While editorial reviews can serve the first use of reviews, that of providing information about the item, e.g. content of the book as here, they cannot typically serve the second use, that of comparing the reviewer's position to the item to the user's position and evaluating the impact of the review from that viewpoint.

Finally, there was certain amount of suspicion about the reviews (Hu, Liu, & Zhang, 2008; Racherla & Friske, 2012), and a large number of positive or negative reviews was considered indicative of the item quality while only one or two reviews, whether positive or negative (although a smaller number of negative ones was necessary for conclusive judgment than positive ones), were viewed with some suspicion. Chen (2008) argues that recommender systems easily lead to herding behavior, and the need for a large number of reviews to lend them credence appears to be one reason for this.

Given the centrality of customer reviews in deciding which book to buy in our study, it can be argued that, while algorithmic recommenders help users by narrowing down the number of items they need to consider, community opinion is essential in making the final decision.

Given that only some reviews are significant to the purchase decision and that one or two reviews were often not considered to provide enough evidence about the quality and content of the book, it is curious that half the participants in our study still used star averages to select items on the list page. At this point, in 2007, the number of reviews was not shown on the list page, so the only cue about eWOM or community opinion on it that was available to the participants was whether or not a book had reviews at all and, if it did, what was the average of the star ratings. On the other hand, the other half selected books based on other factors, such as title, cover picture, and author, checking out books in more detail in complete disregard of the average star rating that the book had.

The presence of eWOM in the form of reviews increases both perceptions of usefulness and social presence (Kumar & Benbasat, 2006) in addition to creating "a sense of community among the frequent shoppers" (Mudambi & Schuff, 2010). Social presence has also been connected to the willingness to contribute in literature (Nov & Ye, 2010). While social presence has not been defined conclusively in literature, we used in our study the definition used in studying Kalas[51]: A perception of "not being alone in the space" (Svensson, Höök, & Cöster, 2005). However, in our study, the arguably rich social texture, features that indicate synchronous or asynchronous presence of others in the environment, of the Amazon interface resulted in only half the participants perceiving Amazon as social. The behavior of those who perceived Amazon as having social presence was clearly affected by this perception. For example, one participant read a review with 95 helpful votes out of 95 votes but did not find it useful. In the end, he decided not to give a negative usefulness vote to it in order not to be a "*killjoy*" when so many others had liked the review.

However, all participants did use social skills and the available social cues to evaluate the position of the reviewer to the items and to themselves to see if the review was meaningful to them. Whether or not an environment is seen as social depends on the individual's definition of social, and what makes one perceive an online environment as having social presence varies even if awareness of other users may still affect their behavior (Leino & Heimonen, 2013). If we had defined social presence as *awareness of others using the space*, we would have had more participants reporting perceiving it. Since it appears that the awareness of others affects behavior (Leino & Heimonen, 2013), this measure might have actually been more meaningful. As it is, we can only conclude that at least some participants perceived Amazon as having social presence and that awareness of others affected the behavior of at least some of the participants. Nov and Ye (2010) have also connected the perception of social presence having behavioral influence in a study of tagging on Flickr.com.

## 10.3 News Recommending: Publication III

III. Juha Leino, Kari-Jouko Räihä, and Sanna Finnberg (2011): All the News That's Fit to Read: Finding and Recommending News Online.
  - What news sources are used today to keep up-to-date
  - How respondents found news online
  - How news is accessed online (mobile vs. computers)
  - Recommending news: Making recommendations and receiving recommendations
    - What types of news is recommended
    - Why news is recommended

[51] Kalas has been described as "one of the most complete social navigation systems ever built" (Riedl & Dourish, 2005) and represents a rare study where a complex information environment was studied as a whole.

    - Role of social media in news recommending
    - Which recommendations get viewed and why
    - Role of recommendation sites, aggregation services, feeds, and recommendation emails
- Who to "follow" for news recommendations
- Implications of current practices
    - News quality vs. social aspects
    - Design implications

As online news has become a major source of news (Purcell et al., 2010) and news recommending has become an important part of online newspapers' strategy in a competition for readers, people face the problem of finding salient news items from the seemingly endless number of items available online (Das et al., 2007; Liu, Dolan, & Pedersen, 2010; Cleger-Tamayo, Fernández-Luna, & Huete, 2012). We set out to understand online news recommending practices as we were designing a news sharing and recommending solution in a research project. Our goal was to understand the underlying dynamics of the news recommending activity taking place in order to be able to design a system that would answer actual user needs and support them instead of imposing new dynamics on users.

In order to collect information about current news recommending activity and news reading online on a large scale, we decided to use a survey approach. While this opens the results to well-known problems connected to all survey studies, e.g. say-do problems and not knowing how seriously respondents took the task, we saw no other practical method for a large-scale study. To reduce the problems, we scanned the responses carefully. Most respondents included verbal comments that indicated that they had given considered responses.

The survey topics were news reading habits (both online and offline), news recommending, receiving news recommending, and impact of recommendations. The survey form was built so that respondents only needed to answer the relevant questions; e.g. if the respondent ticked the *Yes* radio button on being asked if they used news aggregators, the form would open and display more detailed questions concerning news aggregators but if the respondent ticked the *No* radio button, no further questions were displayed on the topic. The survey had 29 questions where different properties (e.g. the importance of the identity of the person making a recommendation) were evaluated on a seven-point interval scale (7 signifying strong agreement) while frequencies were asked using a five-point scale. All these non-open-ended questions had an openable text area for commenting. In addition, there were several open-ended questions[52].

We solicited respondents through mailing lists at three Finnish universities. We raffled four movie tickets among respondents. We received responses from 147 respondents (83 male and 63 female), of whom 58 were students and a further 36 worked at a university. The respondents consequently leaned towards the younger end, with the mode being 20–29 years of age. While this does affect the external validity of the results, for our purposes the sample was suitable, as we wanted to understand the up-and-coming use practices.

#### Contributions in Publication III

The results show that, at least for young adults, online news has bypassed both TV and print newspapers as a source of news. When asked how they *typically* found online news to read, 84% of the respondents reported accessing their favorite news sites directly, 22% using feeds, 20% using news aggregator services (e.g. Ampparit[53] and Google News), and 15% using recommendations systems (e.g. Digg.com). Significantly, 48% also *typically* used recommendations from other people to find news content online.

Roughly, one third of the respondents recommended news at least *Several times a week* and over half at least *Several times a month*. While respondents did not necessarily feel pressure to recommend news items back to people who recommended news items to them, they did emphasize the importance of commenting or otherwise acknowledging recommendations that they received. In effect, reciprocity did not necessarily mean recommending something back but responding somehow to show appreciation.

In fact, much of news recommending activity was aimed at maintaining relations and sharing emotions in addition to sharing information. Recommendations were also often related to on-going conversations taking place online, face-to-face, or both, or meant to function as material for such conversations in the future. In effect, as suggested by Erdelez and Rioux (2000), news recommending is clearly part of overall social behavior and social intercourse between humans; "The act of sharing and receiving information is very personal in all contexts and environments, including electronic ones."

In contrast, when studying recommending online in general, Bernstein et al. (2010) found the strongest motivator for sharing links was to know that the receiver "would appreciate hearing about it". Our results suggest that,

[52] The form was originally in Finnish, but an English translation is available at http://tinyurl.com/6e9n5af.
[53] http://ampparit.com; Ampparit is a popular Finnish-language news aggregation site.

while certainly true, this is more of a superficial reason for recommending, as recommending somebody something we know they like is a social act that strengthens our social ties with them (Ellison, Steinfield, & Lampe, 2007). The underlying motivation to recommend somebody something they would appreciate is likely to maintain social ties although altruism certainly can be a part of it.

When we look at the media that the respondents used for recommending, we notice that social media and IRC are clear winners, with 42% of the respondents recommending news in each. They are closely trailed by instant messaging (IM) that is used by 35% of respondents. Email is used by 30% and *Tell a friend* feature (emailing friends about the item through the system provided by the news site) only by 9%. Twenty-two percent used other means, mostly face-to-face conversations that were used by 14% of the respondents, making it in fact a more common way of recommending news than *Tell a friend*.

Our results appear to contrast with the findings of Bernstein et al. (2010) who found that email was the most common way to recommend content, with 95% using it, while social networking was used by 45% of their respondents (forty people through Mechanical Turkey). However, there are several reasons that may account for the discrepancy. First, news recommending is likely to be different from general content sharing. Second, their categorization is not the same as ours; e.g. we considered Twitter to be part of social media while Bernstein et al. considered it to constitute a whole category by itself. Direct comparisons cannot therefore be made. Also, cultural aspects may be involved; e.g. in our sample of Finns, social media basically means Facebook, as only four respondents out of 147 used Twitter for recommending news [54] while seven respondents out of forty in Bernstein et al.'s (2010) sample used Twitter for recommending content.

In any case, other studies have confirmed the importance of social media in recommending and sharing such content as news, e.g. Gupte et al. (2009) and Lerman and Ghosh (2010). In fact, today social media sites are often "the first to break the important news" (Lerman & Ghosh, 2010). As news recommending is part of normal human social behavior, it is not surprising that today ubiquitous social media plays such an important role in it.

Then again, when we look at which recommendations actually get taken, the importance of social media becomes more debatable. This is an important area of research as we still have only partial understanding

[54] However, since 2010, Twitter has taken off in Finland (see http://www.toninummela.com/blog/2013/07/suomalaiset-liittyvat-twitteriin-nyt.html), and today, the results would likely be different in this regard.

under which conditions and how the selecting of recommended items can be affected (Buder & Schwind, 2012). Overall, many recommendations do get read, as 65% of our respondents claimed to read *All* or *Most* news items recommended to them and 83% claimed to read at least *Half*. There were two important factors that affected whether or not the item was checked out, 1) the sender and 2) the medium. The relationship between the sender and the receiver influenced the decision whether or not to check the recommendation out, which again underlines the aspect of maintaining social relations in news recommending activity.

The second factor, medium, also had a pronounced impact. In effect, only about 62% of news recommendations made in social media were read, while for email, IM, and IRC, the percentage was over seventy, and for *Tell a friend* 90%. When we looked at the text comments, the reason was clear: Respondents regarded recommendations made directly to them as important, and because news recommendations in Facebook are generally not that direct or personal, they are not followed as often. In fact, it seems that other factors, such as title of the recommended item, strongly affect the decision whether or not to follow a recommendation when it is neither direct nor personal.

The findings of a 2013 study by PewResearch Journalism Project support our conclusions. According to it, Americans do not go to Facebook to find news but end up discovering news there; "news is a common but incidental experience" (Mitchell et al., 2013). About half the Facebook users "ever" get news there (Mitchell et al., 2013). When American Facebook users see news recommendations, the topic is the biggest single reason for following the link (70%), trailed far behind by friend's recommendation that only 37% name as a major reason for clicking a news link (Mitchell et al., 2013).

There appear to be two dimensions working here, *quality* (getting good recommendations) and *sociality* (maintaining relations with others). Naturally, the two are not contradictory, as people who know us well are well positioned to recommend news items that are likely to interest us. In direct and personal recommendations, these two work in tandem. However, when sociality is outside of the picture, quality becomes the deciding factor, and then the decision whether or not to read is influenced by quality considerations, such as whether or not the topic is interesting (hence the interest in the title). As social networkers are increasingly becoming strategic posters and propagators of information (Gupte et al., 2009), the links that appear in e.g. Facebook can be aimed at personal image building as much as at propagating an interesting news story.

On average, the respondents in our study who made recommendations used two means for it. Interestingly, these respondents divided into two groups according to the media they used for making news recommendations. Social media, IRC, and IM use correlated as did email, other, and *Tell a friend* use. The two groups are certainly porous and not clearly defined, but some formation is nevertheless evident.

Understanding how news recommending is a part of the larger social intercourse between humans and how it integrates into it allows us to design systems that help integrating news recommending into that larger social intercourse—systems that facilitate the social behavior taking place rather than forcing it to adopt unfamiliar dynamics. Again, if we had simply studied recommender systems and how news recommending takes place in them, we would easily have missed the bigger picture that in fact should guide the building of news recommender systems. Therefore, this study further underlines the importance of studying recommender systems in and developing them with an understanding about the actual use context. Only if we are equipped with an understanding of actual use context can we develop recommender systems that support people in their tasks.

## 10.4 E-LEARNING: PUBLICATIONS IV–VII

IV. Juha Leino (2012a): Case Study: Material Additions, Ratings, and Comments in a Course Setting.
- 2009 version of LSRM (Lecture Slides and Reading Materials)
- Introducing the LSRM system and defining it in learning terms
- Adding additional study materials for the learner community
- Rating materials in LSRM
    - Dishonest ratings
- Commenting materials in LSRM
- Eliciting contributions with compulsoriness
- Social presence
- Evaluation of the LSRM system as an e-learning recommender

V. Juha Leino (2012b): Case Study: Recommending Course Reading Materials in a Small Virtual Learning Community.
- 2009 version of LSRM
- How to approach choosing recommenders for a virtual learning community
- Introducing and discussing the selected features—why these features for this community
- Evaluating the design based on actual use and student perceptions
- Discussion on what worked and what needs to be iterated further

VI. Juha Leino (2013): Recommending Additional Study Materials: Binary Ratings Vis-à-vis Five-star Ratings.

- Description of the LSRM versions in 2009, 2010, and 2011—the versions discussed in this paper
- How students used and perceived the recommending features
- Impact of ratings on selecting items for viewing
- Trust issues
- Five-star rating scale vs. binary rating scale
- Five-star rating scale in e-learning versus other domains
- Characteristics of dishonesty in binary and five-star scales
- Two models that emerged for employing recommenders: Low-cost approach and High-quality approach

VII. Juha Leino and Tomi Heimonen (2013): Improving Evaluation Honesty and User Experience in e-Learning by Increasing Evaluation Cost and Social Presence.
- Description of the LSRM versions in 2011 and 2012—the two versions that are compared in this paper
- Working hypotheses behind the changes in the 2012 version and evaluating how they turned out
- Adding additional study materials
- Selecting materials for viewing
- Evaluating materials and how materials were chosen for evaluation
- Perceived social presence and its impact
    - Social presence of instructors and its impact
- User experience in LSRM
- How to develop the LSRM system further

The four publications on e-learning focus on the additional study material recommender system that we have developed for and used in an undergraduate course on user-centered design (UCD) (lectured annually in the fall semester). As UCD is a vast field, covering several professions, e.g. information architect and interaction designer, covering it comprehensively in 14 hours of lectures is nigh impossible. Consequently, we used to provide students with additional study materials to allow them to learn more about various aspects of UCD. However, this top-down approach has several weaknesses. First, it forces all students to read on the same aspects of UCD, disregarding individual interests. Second, the course instructor is left to decide the suitability of materials for students. While he may be well positioned to determine the quality of the materials, e.g. that they are factually correct and well written, he is not as well positioned to decide the suitability of the material to the students; e.g. whether the materials are too advanced or too basic for the student community. Moreover, it is unlikely that one person can find all the best materials alone.

At the same time, students are already augmenting their course materials at college/university level with online materials and are ready to share these with other students (Hage & Aïmeur, 2008). Consequently, we decided to approach the problem with a bottom-up approach and harness the collective wisdom and efforts of the students in order to let them build a collection of additional study materials for learning more on different aspects of UCD to complement the choice materials that we add. Furthermore, we allowed students to use various recommending features, such as rating, commenting, and tagging (the set of features available varies from year to year, as the system has continuously been developed), to evaluate these materials, including the ones added by the instructors, as the materials are to serve the student community, and therefore the quality and suitability of the materials should be evaluated in terms of that peer community. Also, people have been found to prefer guidance from the people they perceived as similar to themselves (Chen, 2008). Table 4 shows the features that were available to students for evaluating materials in each year.

| | 2009 | 2010 | 2011 | 2012 | 2013 |
|---:|---|---|---|---|---|
| Rating (binary scale) | X | X | | | |
| Rating (five-star scale) | | | X | | |
| Commenting materials | X | X | X | | |
| Tagging | | X | X | X | X |
| Evaluating (five-star rating and commenting coupled) | | | | X | X |
| Commenting evaluations | | | | X | X |
| Collaborative filtering | | | | | X |

**Table 4**. The recommending features available for evaluating/annotating materials in each year.

The system was implemented as a web page on the course web site (requires logging in) using HTML and JavaScript to implement the interface, AJAX to connect to the server without having to reload the page, and PHP to interface with the Postgresql database. We call the system LSRM (Lecture Slides and Reading Materials), as the page gives students access both to lecture slides and additional study materials.

Recommending took place at two levels in the system, 1) adding links to materials that one had found useful, and 2) evaluating and annotating the added materials with various recommender features. Adding a link to a system is an implicit recommendation (Varian & Resnick, 1997; Lerman, 2007), as the act of adding a material to the repository implies that the person adding it has considered it useful for the community members.

We decided to use annotation approach because algorithmic approach would not have had sufficient time or amount of data to work; e.g. the collaborative filtering recommender system in Kalas failed to be properly bootstrapped in spite of the fact that the system was used for half a year by 302 users (Svensson, Höök, & Cöster, 2005). Our time frame was limited to one semester, about three-and-half months, and the number of students was never over 55 and the number of materials never over 124. Consequently, we saw no other option than to use a non-algorithmic approach based on presenting the community opinion rather than a list of recommended items. Moreover, non-algorithmic approach provided value to the community immediately and allowed us to let the student who made a contribution to immediately see their contribution to the community, which has been shown to have positive social outcomes (Rashid et al., 2006).

In effect, LSRM is similar to Tapestry in that it also functions as a repository and not only as a filtering mechanism (Goldberg et al., 1992). Also, LSRM is very similar to Digg.com in that the basic functionality in the two is the same: Users submit links to materials that they have found online and other users evaluate those materials (Lerman, 2007).

Another factor affecting our design decision was that in literature, it is suggested that with small communities, it can be useful to display the individual ratings and reviews of community members so that each can draw their own conclusion about the strength of a recommendation (Schafer, Konstan, & Riedl, 2001). Reviews in any case are not entirely machine-readable, and the text comment can provide "an understanding of why a particular item should be favored or disfavored, and comments may be the only way to help a customer navigate through substantial disagreement among people who have agreed before" (Schafer, Konstan, & Riedl, 2001).

In the 2009 version of the system, rating (binary scale; *Useful* or *Not useful*) and commenting on an item were decoupled; a student could rate an item without commenting it, or comment without rating an item (one could not rate an item one had added but could comment to enable discussing). The idea was to keep the rating effort as low as possible to encourage rating. The design reasoning of the 2009 version is explained in detail in Publication V that focuses on the design decisions of LSRM and provides a proof of concept for the design, as well as discussing ideas on how to improve it in the next iteration. In effect, evaluating "the quality of documents or information in general" is "[o]ne of the outstanding problems in information processing" (Lerman, 2007), and so designing a system to support this in not trivial.

Moreover, "[f]inding a 'good' paper isn't a trivial task" but "a multiple-step process that typically entails the users navigating the paper collections, understanding the recommended items, seeing what other users like and dislike, and making decisions" (Tang & McCalla, 2009). From the start, our purpose for using LSRM on the course was two-fold: 1) to encourage students to read more widely on UCD and to develop a habit of following online sources for current topics in the field, and 2) help students develop their information literacy skills, as finding, evaluating, and using information from various sources is seen as a survival skill in this information-intensive age and as crucial for professional success after graduation (Kiliç-Çakmak, 2010). In effect, both finding high-quality online study materials (be they columns, blogs, scientific papers, presentations, or videos), determining their quality and suitability for the community, and evaluating the materials others had added all allow students to hone their information literacy skills. Also, as their additions are evaluated by the community, they get feedback on their performance, further allowing them to improve these skills.

| | **2009** | **2010–2012** |
|---|---|---|
| Design assignment | 80% | 70% |
| Smaller assignments (10 pieces) | 20% | 20% |
| Online activity (LSRM) | Max. -10% | 10% |
| Attendance (extra) | 10% | 10% |
| ***Max. total*** | ***110%*** | ***110%*** |

**Table 5**. Course grading.

In 2008, we basically had the same system running as in 2009 but adding materials and evaluating them had been voluntary, i.e. it did not affect the student's grade. Reading materials was and has continuously been voluntary in the sense that students are not examined on the materials. As there were only a few contributions in 2008, we decided to use a stick to encourage contributions: Students were required to contribute one material link and five ratings—commenting was voluntary—or they faced a punishment of 10% off their grade. In 2010, the stick was turned into a carrot, as we moved 10% of the course grade to online work; now students were required to add two material links to LSRM and rate five[55]. Commenting was still voluntary, as was tagging, a feature that was added to the 2010 version. Table 5 shows the grading of the course in 2009–2012. We will discuss using compulsoriness to encourage contributing in e-learning below in more detail.

[55] According to *loss aversion*, introduced by Kahneman and Tversky (1979), people prefer avoiding loss to acquiring gains, meaning that we feel loses more strongly. In 2009, we therefore employed the fear of loss to motivate students (i.e. stick) while afterward, we made active participation appear as a chance to gain credit points (i.e. carrot). Still, it is a question of framing, as in both cases, the effect was fundamentally the same.

In the 2010 version, tagging was added to allow mini-reviewing of items. Some students had felt in 2009 that ratings should be accompanied with a comment to justify the rating and to show the thinking behind it. Also, some students felt that the person adding a link should have told why they had added it. Tagging was added as a low-cost approach to opinion-giving, since it has been shown to be used like that in any case (Lee, Son, & Han, 2007). Tagging was implemented as a separate feature—anybody could tag a material, including the adder, but tagging was not required. As tagging was little used in 2010, in 2011 tagging a material one added was made a compulsory part of adding a material. However, other students could also tag materials others had added.

One observation of the use in 2009–2010 was the presence of large-spread dishonesty in ratings: Students rated items without actually opening them. (We discuss dishonesty and reducing it through design below in more detail.) While the actual amount of dishonesty and student perceptions of the amount of dishonesty were not big enough to cripple the usefulness of the system, as discussed in Publication IV, we decided to try to reduce it by increasing the complexity and cost of rating effort. Therefore, in 2011, instead of a binary rating scale, we employed a five-star rating scale. Another reason for the change was to allow students to make more fine-grained evaluations, as people have been found ready to exert more effort if they are given a chance to be more expressive (Cremonesi, Garzotto, & Turrin, 2012b; Pommeranz et al., 2012). Moreover, using a more fine-grained evaluation was expected to increase information literacy benefits, and a five-star interface provided a familiar way to visualize the average of the ratings for a material link. Also, the five-star scale has repeatedly been found to be popular among users (Cosley et al., 2003; Gena et al., 2011; Sparling & Sen, 2011).

As using the five-star rating scale almost halved the dishonesty in ratings, we decided to try more of the same, and combined rating and commenting into an evaluation so that users had to provide a title, rating on a five-star scale, and a comment as an evaluation in 2012. We also added a possibility of commenting evaluations to allow exchanging opinions without evaluating the item, as otherwise those who already had evaluated an item would not be able to take part in the discussion. Commenting was voluntary in the sense that it had no impact on the grade.

Also, as all contributions had been anonymous this far, we decided to give students individual presence and a possibility of reputation formation by having them adopt a nickname for the system. Tagging items added by others was still anonymous but adding materials, evaluating them, and commenting evaluations were not. The reason for using nicknames was to increase, facilitate, and enhance sociality that was already present in the system and to increase honesty through social presence and pressure.

After we look at LSRM in educational terms and the data collecting for the four e-learning publications, we look at each paper briefly to outline its content, and then discuss the contributions that these studies make to understanding human factors in recommender use in e-learning. As the papers build on each other and the system has been iterated based on the understanding gained from the previous study, it makes more sense to discuss the main contributions of these studies as blocks rather than publication by publication.

#### LSRM in Educational Terms

As discussed, e-learning can be divided roughly into *formal* and *informal* learning (Drachsler, Hummel, & Koper, 2009; Manouselis et al., 2011). *Top-down* approach is commonly seen as suitable to formal e-learning, as structure and learning materials can be provided by domain professionals whereas *bottom-up* approach is seen as more suitable to informal learning, as learners have to contribute the information, e.g. ratings and tags, on which the recommendations are based (Drachsler, Hummel, & Koper, 2009; Santos & Boticario, 2010).

LSRM in a sense challenges this clear-cut division, as the context, a university course, is definitely within the realm of formal learning but, at the same time, the approach to recommending was very much bottom-up, as students provided about half the materials and most evaluations—and the materials or evaluations by the instructors were not given any special status; the instructors were simply members of the community as far as LSRM was concerned. Also, the fact that students were not tested on the materials leans towards informal learning—students had to take responsibility for their own learning.

LSRM can also be characterized as a learning network, as the students using it constituted a closed community of peers, its recommendations were largely based on student contributions, and it used Web 2.0 approaches and technologies as enablers (Manouselis et al., 2011). Also, learners in learning networks typically need support in finding and selecting suitable sources of information (Drachsler, Hummel, & Koper, 2009), something that LSRM, by its very nature, was designed to support and facilitate.

Overall, the course setting presents *blended learning*, as it combined online and face-to-face learning. Blended learning is part of the larger shift in education away from teacher-centered learning towards learner-centered learning and conceptually from teaching to learning (López-Pérez, Pérez-López, & Rodríguez-Ariza, 2011; Ruohotie & Nokelainen, 2003). Blended learning is seen as providing more student autonomy and flexibility, resulting in a deeper understanding of the subject material and capability for reflection (López-Pérez, Pérez-López, & Rodríguez-Ariza, 2011). Blended learning is seen as very much the direction that learning will take (Garrison & Kanuka, 2004), which underlines the importance of studies on e-learning to guide the development.

#### Data Collecting

The LSRM web page records every meaningful click on it to the database, providing a virtual click-by-click picture of the activity on the page. Our behavioral data is based on this use-log data. Our subjective response data is collected by a questionnaire that, on average, more than half the students provided (Table 6). In 2009, students were told to fill in the online questionnaire as part of the course activity, and they were provided with a written summary of the feedback that they had given. Since then, providing feedback has been purely voluntary and movie tickets have been raffled among those who have provided feedback. In 2012, in addition to the questionnaire data, three students were also interviewed. Because the author here, the lecturer of the course, had designed the system and students knew this, the interviews were conducted by the teaching assistant, who is also the co-author in Publication VII that focuses on the 2012 data.

| | **Students on the course** | | **Questionnaire respondents** | |
|---|---|---|---|---|
| | *No. of students* | *No. of females* | *No. of students* | *No. of females* |
| *2009* | 32 | 10 (31%) | 23 (72%) | 7 (30%) |
| *2010* | 55 | 14 (25%) | 22 (40%) | 8 (36%) |
| *2011* | 37 | 8 (22%) | 19 (51%) | 5 (26%) |
| *2012* | 36 | 18 (50%) | 20 (56%) | 12 (60%) |
| ***Total*** | ***150*** | ***50 (33%)*** | ***84 (56%)*** | ***32 (38%)*** |

**Table 6**. Number of students who provided feedback on LSRM via the online questionnaire.

#### A Brief Glance at Publications and Their Contributions

***Publication IV*** focuses on the 2009 experiences of the system. It considers using compulsoriness to elicit student contributions in the context of e-learning; it is the first work that we know of that discusses dishonesty in e-learning and how it differs from dishonesty in e-commerce; considers the impact of social presence on contributing and student behavior in e-learning; and provides the field with experiences in employing recommenders in e-learning with actual users and real use context, something that is needed as most systems are, at best, at a prototype level and very few systems have been evaluated with users (Manouselis et al., 2011).

***Publication V*** discusses the design and design choices in LSRM and evaluates the 2009 design against the student feedback, providing a proof of concept for the system. It discusses what worked and why and what needs to be re-worked on in the next iteration of the system. The publication provides a deep discussion on designing systems to help students use online materials as a virtual learning community within the framework of formal learning. As mentioned, designing such systems is not trivial (Lerman, 2007; Tang & McCalla, 2009).

***Publication VI*** focuses on 2011 experiences of moving to a five-star rating scale and contrasts these experiences against the 2009 and 2010 experiences with a binary rating scale—the simple changing of the rating scale almost halved dishonesty. Also, as the 2009 and the 2010 versions were very similar and produced similar user experience both in terms of subjective responses and use-log data, we believe that differences are due to the changes in the recommender feature and not due to a different group of learners responding differently to the interface. As five-star rating distributions in e-commerce have raised questions about how well ratings can actually reflect actual item quality (Hu, Pavlou, & Zhang, 2006; Talwar, Jurca, & Faltings, 2007; Hu, Zhang, & Pavlou, 2009), we contrast our experiences in the domain of e-learning to those reported in literature in the domain of e-commerce. To our knowledge, this has not been done before, as a lack of actual use experiences in e-learning has prevented it. Finally, we also discuss two models for using recommender features in e-learning based on our experiences over the three years.

***Publication VII*** focuses on 2012 experiences when the cost and complexity of evaluating materials was raised by requiring an evaluation that combined a title, rating on a five-star scale, and text comment, and anonymity was replaced with nicknames that provided students with an individual presence in the system. Increased perceived social presence and cost of evaluating resulted in 100% honesty and increased perceptions of learning due to using the system. The 2012 study—and version of the system—show conclusively how to combat dishonesty in a formal e-learning setting. Also, we discuss the implications of social presence to such e-learning systems.

#### Using Compulsoriness to Encourage Contributing in e-Learning

As we noticed in 2008 when contributing was voluntary, under-contribution is as serious a problem also in e-learning as it is in other domains (Recker, Walker, & Lawless, 2003; Ling et al., 2005; Farzan & Brusilovsky, 2006). In fact, "[e]ncouraging users to contribute has become one of the most important challenges" in the field of e-learning (Farzan & Brusilovsky, 2006). As formal e-learning context allowed us to use compulsoriness, we solved the under-contribution challenge with it, as also Recker, Walker, and Lawless (2003) had done. Interestingly, moving from the stick (punishment of 10% for not making the defined minimum contribution) to the carrot (online work constituting 10% of the grade) did not change the user experience of the system in terms of students' perceptions or use-log data, so it appears that both approaches can be used.

In addition to general under-contributing problems in online communities, it is important to get larger fractions of communities to contribute, as today in e-commerce it is the opinionated users who voice their opinions while those with moderate outlooks do not bother to voice theirs, leading to serious questions concerning how well eWOM actually reflects the real item quality (Hu, Pavlou, & Zhang, 2006; Ling et al., 2005; Talwar, Jurca, & Faltings, 2007; Hu, Zhang, & Pavlou, 2009). In effect, in e-commerce, rating distributions tend to be J-shaped as most ratings are very positive, some are very negative, and there is precious little between (Kadet, 2007; Hu, Zhang, & Pavlou, 2009).

Interestingly, in e-learning, the use of compulsoriness appears to solve this problem, as it results in a larger number of the community members to voice their opinions. When we looked at the distributions of ratings in 2011 and 2012—when a five-star scale was used—we noticed that they were much closer to normal distribution (although statistically not normally distributed) than rating distributions reported in literature concerning e-commerce. In effect, compulsoriness appears to democratize eWOM by making a large fraction of the community voice its opinion. Another reason appears to be that, while in e-commerce love-it-or-hate-it models appear to have explanatory power, in e-learning emotions are not that burning, which appears to allow for more level-headed evaluations of the items. In effect, ratings in e-learning may reflect the actual item quality better than they do in e-commerce.

Naturally, compulsoriness has to be used judiciously (Leino, 2012a), as asking too much easily leads to demotivation (Ling et al., 2005). However, reasonable amount of compulsoriness that has clear learning benefits is not resented at least in the context of formal e-learning (Leino, 2012a). Moreover, setting reasonable and specific, clear goals has also been shown to motivate contributions outside of e-learning setting (Ling et al., 2005).

Overall, use of compulsoriness in e-learning has received little coverage in e-learning literature and its impact on rating patterns has not been discussed elsewhere to our knowledge. Also, we are the first to compare rating distributions in e-learning to those in e-commerce.

#### Dishonesty and Reducing It in e-Learning

However, use of compulsoriness has the downside of easily increasing dishonesty. Dishonesty here is defined as evaluating a material without having viewed it—we cannot know if students had read materials before rating them but LSRM allows us to know if they had opened the material, and therefore we use the term *viewing*. Commenting, for example, has been voluntary throughout the use of LSRM and there have been no dishonest comments. In contrast, the moment rating items was made compulsory, some students took the easy way and clicked away ratings without bothering to even open the materials they were rating.

Clearly, dishonesty is motivated differently in e-learning in comparison to e-commerce (Leino, 2013): In e-commerce, the goal is to either promote an item to get others to buy it or to demote an item to prevent others from buying it (e.g. in the case of a competing product) (Herlocker et al., 2004; Lam, Frankowski, & Riedl, 2006) while in e-learning, it is more about trying to get credit points without earning them due to lack of time or laziness, as some students professed (Leino, 2012a).

In effect, dishonest ratings in LSRM may have had a relatively small distorting effect on the accuracy of the system, as most dishonest raters simply followed the herd and rated items that have been rated before and rated them in accordance with earlier raters. In effect, different motivations for dishonesty in e-commerce and e-learning result in different approach to dishonesty—following the existing ratings is not a very effective approach to shilling, to say the least—and it means that tackling dishonesty and recognizing dishonest ratings need to use different approaches in different domains.

In 2009 and 2010, dishonest ratings constituted about 40% of the ratings. Telling students in the interface that they should read the material before rating it would likely only teach students to game the system by opening the material and then rating it without actually having read the material. Consequently, we decided to tackle the problem by making the rating decision more complex in the 2011 version to make cheaters give up their cheating ways. In effect, by simply changing the rating scale from a binary one to a five-star one reduced dishonesty to 26% of the ratings. However, the cost involved was that the number of students not rating materials increased and the number of students rating more than required number decreased, resulting in a statistically significant decrease in the number of rated materials. This phenomenon of increasing rating scale granularity resulting in fewer ratings has also been observed elsewhere (Sparling & Sen, 2011).

However, we believed that increasing the cost and complexity of evaluating materials even higher should increase honesty even further. We also hoped that student perceptions of hard work bringing learning benefits would offset the tendency of increased effort leading to fewer ratings. Consequently, in 2012, we required students to evaluate materials by providing a title, rating on a five-star scale, and text comment instead of simply rating an item. In addition, we gave up anonymity and required students to use nicknames for increased social presence and to give them individual presence and therefore reputation in the system. The result was 100% honesty—all students viewed all items that they rated before rating them and the average time between opening of a material and rating it almost doubled—without the average number of items rated per student decreasing in comparison to 2011. Importantly, the 2012 version not only resulted in 100% honesty but also in higher perception of honesty, as no student questioned how carefully others had made evaluations, while in 2011, in comparison, there was some suspicion that some students might have clicked ratings without actually reading the materials or not reading it carefully. (Leino & Heimonen, 2013)

Moreover, in 2012 students viewed more than 2.5 times more materials than in 2011. In effect, students evaluated much fewer materials in proportion to how many they read in 2012 than in 2011. The reason is at least in part that students felt that they needed to have something significant to say about a material to evaluate it, and if they had nothing new to add to the earlier comments of the material or they did not have a clear opinion of the material, they did not evaluate it. (Leino & Heimonen, 2013)

In effect, increasing rating cost and complexity had a very positive effect and proved to increase honesty significantly and, as is discussed below, resulted in increased perceptions of learning information literacy skills in addition to perceptions of having learned more about the topic of the course. Consequently, the design guideline by Pu, Chen, and Hu (2012) to "[m]inimize preference elicitation in profile initialization" does not apply to all situations at least in e-learning; longer and more complex elicitation can in fact result in a better user experience. In fact, as discussed, minimizing effort in preference elicitation together with compulsoriness had a detrimental impact by increasing dishonesty.

Then again, the guideline may not apply in other domains uniformly, either, as asking more questions appears to have only a marginal effect on perceived enjoyment, users are ready to exert more effort if it enables them to be more expressive, and as long as the higher effort leads to significant improvements in recommendations, the global satisfaction is not affected (Gretzel & Fesenmaier, 2006; Cremonesi, Garzotto, & Turrin, 2012b). In effect, in recommender systems, there appear to be few hard and fast guidelines that apply in all domains and contexts.

Also, while Recker, Walker, and Lawless (2003) concluded from their experiences that in e-learning "simple review schemes work best" and that "low-overhead means of collecting review data are preferable", leading them to consider using implicit preference collecting, our results point to the opposite direction: By making evaluating items harder we improved the overall experience and honesty. In any case, while implicit data can lead to good profiling, much more of it is needed because of its noisiness (Gadanho & Lhuillier, 2007; Konstan & Riedl, 2012), and because of lacking accuracy, it can damage the system's precision (Farzan & Brusilovsky, 2006). In our context, collecting implicit data would not be feasible due to the low number of interactions and short time period, and the same applies to many e-learning settings.

#### Improving Learning through Recommenders

Recommender systems have been suggested to offer learning benefits beyond finding suitable learning objects. First, providing ratings and reviews on objects is seen as requiring reflection and leading to deeper elaboration and learning (Recker, Walker, & Lawless, 2003; Buder & Schwind, 2012). In effect, explicit elicitations of user opinions "appears to be a promising approach for educational recommender systems" (Buder & Schwind, 2012). Consequently, contributing opinions on items should be seen as a learning opportunity rather than as a burden (Buder & Schwind, 2012). Also, Recker, Walker, & Lawless (2003) believed that reviewing web resources should lead to gains in information literacy skills.

Our results provide evidence for these suggestions. Both having to find study materials to add and to evaluate added materials were seen as leading to more careful reading of materials already in the 2009 version of LSRM. However, while Recker, Walker, and Lawless (2003) concluded that "simple review schemes work best", our data leads us to conclude that, as long as the effort is in proportion to the benefits gained from the effort and to the workload of the course, extensive evaluating of objects leads to better results. As discussed, more extensive evaluating improved both honesty and perceptions of honesty. This in turn led to students feeling social pressure to do right by other students, resulting in a positive user experience. Significantly, the more extensive the required evaluating, the higher the number of students who mentioned having gained clear learning benefits from using the system. Students discussed both benefits related to learning about the topic of the course (UCD) and related to information literacy. In fact, only with the 2012 version, a larger number of students mentioned information literacy skill benefits. In effect, while students had reported learning benefits from the 2009 version onward, they were mentioned increasingly more frequently as the cost and complexity of evaluating items also increased, leading us to conclude that increased cost and complexity of evaluating items does result in increasing learning benefits and increasing awareness of the said benefits. As Buder and Schwind (2012) suggest, "educational recommender systems should involve learners by challenging them."

In effect, our results attest to recommender systems having learning benefits beyond recommendations of suitable objects and that they can be used to facilitate learning itself, as they can be used to involve and engage students to take active part in the constructive process of learning, as suggested by Buder and Schwind (2012). However, this also means that simply stating that "[r]ecommender systems are decision-support systems" (Svensson, Höök, & Cöster, 2005), "designed to help the user make better choices from large content catalogs" (Knijnenburg et al., 2012), makes recommender systems appear simpler and less powerful than they really are. Recommenders have, in fact, a much greater potential than simply helping us make decisions concerning objects in e-learning (Buder & Schwind, 2012) and arguably also in other domains. However, recommender systems have to be adapted to different domains so that their potential in those domains comes to full fruition; without such adaptation, recommenders would miss their potential of becoming potent means for learning and simply guide us to various items that are likely to interest us and help us choose between those items.

#### Enhancing User Experience and Social Presence through Recommenders

It is increasingly recognized that user experience is the decisive factor in recommender system adoption and continued use (McNee, Riedl, & Konstan, 2006a; Konstan & Riedl, 2012). However, in addition, the use of recommenders should enhance the user experience of the environment where they are embedded by turning the "experience into a social and pleasurable one" (Svensson, Höök, & Cöster, 2005; see also Herlocker et al., 2004; McNee, Riedl, & Konstan, 2006a). Herlocker et al. (2004) list *Just browsing* as one user task typical to recommenders in e-commerce and other domains, stating that users find recommenders "pleasant to browse" and use such recommender sites as Amazon.com and MovieLens "even when they have no purchase imminent." They further state that "a substantial use of recommenders is simply using them without an ulterior motive" (Herlocker et al., 2004). Other studies have also found that users value and like recommendations (Recker, Walker, & Lawless, 2003) and that recommenders "inspire people and help them explore the space", thus enhancing the experience and making using an environment more pleasurable (Svensson, Höök, & Cöster, 2005).

In fact, it goes almost without saying that recommenders in e-learning should also enhance the user experience of the learning environment. In the case of LSRM, the 2009 version already showed that recommenders both increased sociality and enhanced the user experience, as students stated that the ratings and comments made the otherwise "*lifeless list of links*" more "*living*" and "*interesting*". However, while the student feedback continuously shows that the recommender features enhanced the experience, it is in the 2012 version that this really shines through. There appear to be two main reasons for this. First, costly enough evaluating removed suspicions that others were taking the easy way and just clicked ratings without reading materials, resulting in a virtuous circle of the positive experience leading to diligent work that in turn enhanced the experience. Second, the use of nicknames led both to increased perceptions of social presence and social pressure, as one's evaluations and materials carried one's nickname. This social aspect also enhanced honesty and made students approach the work with a good attitude, contributing further to the virtuous circle.

However, it is important to remember that already in 2009 some students had reported having perceptions of social presence. In effect, adding nicknames did not introduce social presence to the system even though nicknames enhanced perception of it significantly. Therefore, simply having recommender features annotating study materials with community opinion was enough to engender some perceptions of social presence.

Engendering social presence, or "[m]aking the community visible", is important as it can fuel collaborative spirit that results in more and higher-quality contributions (Buder & Schwind, 2012). In fact, when motivating factors for eWOM in e-commerce have been studied, Hennig-Thurau et al. (2004) found that desire for social interaction, concern for other consumers, and potential to enhance own self-worth were significant motivations in addition to desire for economic incentives, while Cheung and Lee (2012) found similarly that sense of belonging to the community and enjoyment of helping others were the most critical factors while reputational factors were only marginally significant. Herlocker et al. (2004) also mention expressing self and helping others as major motivations in addition to improving one's profile to get better recommendations, in a sense an economic motivation.

In effect, central motivators are helping others, social interaction (including belonging), self-expression, and enhancing one's own reputation in addition to economic incentives. Economic incentives in e-learning tend to center around the grade and credit points (Recker, Walker, & Lawless, 2003), as is also the case with LSRM. The other motivators, however, are something that can easily be provided in e-learning and that LSRM in fact provides. The 2012 version also provided a vehicle for enhancing one's reputation with one's work as nicknames were used as well as a chance to socially interact and help others to find useful study materials. In fact, some students clearly commented that the fact that their nickname was next to their contributions made them work harder on them.

All these factors have popped up in student questionnaire responses as motivating factors. Altruistic motivations appear especially significant.

However, even the 2012 version of LSRM failed to engender 100% participation; out of 32 students, seven added no evaluations and eight no materials. Others added at least the required amount, with four evaluating more than five materials and 11 adding more than two materials.

Consequently, in 2013, we are interested to study student motivations behind contributing and not contributing. One such possible factor is knowledge self-efficacy, i.e. one's beliefs about one's abilities to contribute (Cheung & Lee, 2012). Although Cheung and Lee (2012) found no evidence that self-efficacy had any significant impact on contributing eWOM in online consumer opinion platforms, some of our students have expressed concerns about their abilities to contribute meaningfully, e.g. "*You saw what others had praised so you read it with interest, too, but I myself didn't like to evaluate. Especially since I'm not a professional or somebody who knows a lot, so my comments may have appeared pretty bad for somebody who knew more*" (Leino & Heimonen, 2013). While such statements have been few, it appears that the fact that LSRM is a closed community of peers is an important factor to its success, as for example two out of the three students interviewed in 2012 said that it was much easier to contribute in LSRM than it would have been on the Internet, as there are so much more knowledgeable people on the Internet. Consequently, knowledge self-efficacy factors may turn out to be more significant here than in the consumer opinion platforms.

In any case, it appears that the fact that LSRM is a small, closed community of peers creates a safe environment that encourages participating and contributing, and this is something to keep in mind when adding recommender systems to e-learning environments.

Social presence is also important, as it engenders positive behavior. Not only did it activate students and enhance the experience of the environment, but it also made students more aware that what they did affected others, resulting in them wanting to do their share with due diligence, e.g. "*Sociality in the service affected my actions significantly: It affected so that I wanted to select as suitable articles as possible to add to the page and that I wanted to say something more deep than just 'quite nice' in the evaluation*" (Leino & Heimonen, 2013). Interestingly, even students who claimed not to perceive the environment as having social presence mentioned that awareness of other users did affect their behavior, e.g. "*I didn't really feel others to be present that much. Still, the thought that others see what I add there affected what materials I added and what kind of evaluations I made. I.e., I did my job with care*" (Leino & Heimonen, 2013). This is significant, as Konstan and Riedl (2012) call for research "to understand how recommenders can be tuned to achieve positive social outcomes". In effect, within e-learning, increasing social presence and awareness of others through social cues that recommenders can provide is one way to engender positive social behavior.

#### Recommender Impact on Item Selecting

While students using LSRM have largely been happy with the system, in the earlier versions of LSRM they have emphasized in their comments the system's ability to guide them to better materials and to help them avoid the bad ones. However, with the 2012 version, students began to emphasize much more the learning benefits they gained from using the system.

Which materials students view is affected both by volume and valence in addition to other factors, such as title, as also in e-commerce (Chen, 2008; Cui, Lui, & Guo, 2012). Just like consumers are not interested in positive reviews of products they do not intend to buy in the first place (Racherla & Friske, 2012), students in LSRM would not open links the topic of which did not interest them no matter how many positive ratings they had gotten.

If the topic was interesting, then the ratings became more significant. Interestingly, even in 2012 when ratings were perceived as honest, the number of ratings correlated better with the number of students who viewed the item, $r(103) = .64$, $p = .01$, than the star-rating average, $r(103) = .52$, $p = .01$. While the result could be construed to mean that perhaps some other factor, e.g. material title, attracted students to certain materials and these materials therefore ended up getting more evaluations, student comments point out that both the valence and volume were significant heuristics in selecting materials. (Leino & Heimonen, 2013)

According to student comments, negative comments especially had a strong impact, as has also been observed in e-commerce (Berger, Sorensen, & Rasmussen, 2010; Cui, Lui, & Guo, 2012; Racherla & Friske, 2012). At the same time, however, some students admitted that a negative evaluation made them open the link to see for themselves why the material was considered bad.

For many, it was the existence of ratings that attracted them because that meant that they could compare their experience to that of the previous rater(s), i.e. there were social undertones to selecting already-rated materials. At the same time, some students purposely wanted to read materials that had no evaluations so that their evaluations would be useful for the community; i.e. there were also altruistic aspects to how materials were selected for viewing and here previous evaluations functioned as a negative signpost.

In summary, while both valence and volume of the ratings for items functioned as heuristics for selecting items, their impact was not always straightforward, and there were many social, including altruistic, factors that affected the equation. Predicting which items get read requires a multi-perspective approach, and ratings are just a piece of the puzzle, albeit an important piece.

# 11 Conclusion

Today, user-centric factors are widely considered crucial to recommender systems (e.g. Herlocker et al., 2004; McNee, Riedl, & Konstan, 2006a, Redpath et al., 2010; Konstan & Riedl, 2012; Cremonesi, Garzotto, & Turrin, 2012a; Pu, Chen, & Hu, 2012). However, studying and understanding these “more complex and articulated to operationalize” factors (Cremonesi et al., 2011) has been neglected to the detriment of the field (McNee, Riedl, & Konstan, 2006a), in part because studying them is very challenging, expensive, and resource-demanding (Kumar & Benbasat, 2006; Redpath et al., 2010; Cremonesi et al., 2011; Knijnenburg, Willemsen, & Kobsa, 2011; Buder & Schwind, 2012; Konstan & Riedl, 2012; Cremonesi, Garzotto, & Turrin, 2013), as it requires “developing a system, including both algorithms and user interface, and carrying out field studies with long-term users of the system—the only reliable way of measuring behavior in a natural context” (Konstan & Riedl, 2012).

Instead, at least until recently, research efforts have largely focused on system-centric factors (e.g. Herlocker et al., 2004; Cremonesi et al., 2011; Knijnenburg, Willemsen, & Kobsa, 2011; Buder & Schwind, 2012; Knijnenburg et al., 2012; Pu, Chen, & Hu, 2012). Consequently, while the wide adoption of recommender systems has made being able to characterize user experience and users’ attitudes towards recommender systems crucial (Pu, Chen, & Hu, 2011), there are many, many open questions concerning user-centric factors in the field of recommender systems (Buder & Schwind, 2012; Konstan & Riedl, 2012; Racherla & Friske, 2012).

As recommender systems and approaches cannot be directly transferred between different domains (e.g. Tang & McCalla, 2004a; Drachsler, Hummel, & Koper, 2009; Ghauth & Abdullah, 2010; Santos & Boticario, 2010; Buder & Schwind, 2012)—and not necessarily even within a domain (Ghauth & Abdullah, 2010; Granić & Adams, 2010)—understanding user factors and user experience necessitates studies in each domain. In the past, we have been too eager to apply experiences from one domain to another, leading e.g. to guidelines that state that in designing recommenders, the preference elicitation should be made easy to the user based on e-commerce experiences (Pu, Chen, & Hu, 2012) when e.g. in e-learning, collecting explicit preferences actually constitutes a learning opportunity and viewing it as a burden is plain wrong (Buder & Schwind, 2012).

Moreover, many studies focus on recommender systems or on their parts (e.g. algorithms, recommendation list diversity, or explanation interfaces) although recommender systems are typically employed as parts of a larger complex information ecosystem; e.g. Amazon.com is not a recommender system but a complex e-commerce environment that employs many kinds of recommender systems. While the studies focusing on various aspects of recommender systems are certainly important, we also need to complement them by studying user factors and experiences in larger wholes to understand how recommender systems affect them and how to improve the contribution of recommender systems in turning a *space* into a *place*, in terms of Harrison and Dourish (1996). Complex environments bring forth emergent behavior that cannot be traced back to the actions of any individual user (Lerman, 2007; Buder & Schwind, 2012), and studying what actually happens in such environments and what kinds of user strategies emerge to deal with the complexity can only be done by studying such environments when they are used by actual users (Lerman, 2007).

The case studies discussed in the Publications I–II and IV–VII offer us windows into the actual behavior, motivations, and strategies of real users in e-commerce and e-learning. Especially in e-learning, an inchoate domain where recommender systems are increasingly considered, our case studies provide direly needed experiences of actual use by genuine users (Manouselis et al., 2011) that can guide implementation efforts. Also, as experiences are few, we have been in a position to start discussion on dishonesty in e-learning.

In e-commerce, there are more studies of how actual users use such complex ecosystems as Amazon.com. However, many are based on data collected from such services (e.g. such data as star-rating, reviews, and sales position of an item have been used in various ways to study the impact of ratings and reviews on sales) or on questionnaire data, and studies that combine observation with interviews and questionnaires to reduce the effect of the say-do problem are still not many. In effect, even in e-commerce, our study, although already conducted in 2007, still affords a rare glimpse into the use strategies that have emerged in such environments; how users use and behave in such environments and what motivates that behavior.

Finally, Publication III focuses on prevailing news recommending practices and habits, affording a look at how news is found and read today. In laying bare the underlying dynamics of news recommending activity and showing how it connects to the wider social behavior and social intercourse between people, it provides us with a means to approach designing news recommending applications that support people rather than force them to adapt their practices to the tools.

At the same time, however, case studies can only function as windows. All case studies of actual use are subject to contextual influences and confounding factors, and to claim that the results from any one case study represent any kind of a universal truth is, it goes without saying, rather silly. When it comes to case studies, we need many case studies and when their results start to point to one direction, we have something that we can start to generalize but even then with proper caution. Nevertheless, our results provide material for many interesting hypotheses that can be studies in more controlled experiments, in addition to providing experiences that can be reflected against experiences that similar studies will contribute.

In closing, we would like to underline the importance of also studying recommender systems as parts of complex information environments under actual use by genuine users in order to understand how to build them, so that they serve their users and allow their users to reach their goals—and not just accomplish tasks—in a pleasant way. German military strategist Helmuth von Moltke said that, “No battle plan ever survives contact with the enemy”[56], and, in the same way, our best laid designs often go awry when they come into contact with actual users in actual use contexts. Consequently, we need to study what happens in the complex information environments and how users experience them to make them better step by step. In the same sense that von Moltke said that, “The tactical result of an engagement forms the base for new strategic decisions because victory or defeat in a battle changes the situation to such a degree that no human acumen is able to see beyond the first battle”, it is unlikely that we get a complex information environment perfect right off the bat.

[56] http://en.wikiquote.org/wiki/Helmuth_von_Moltke_the_Elder—the second quote by von Moltke is also from here.

# 12 References

Adomavicius, G., Bockstedt, J., Curley, S., & Zhang, J. (2013). Do recommender systems manipulate consumer preferences? A study of anchoring effects. *SSRN*. doi:10.2139/ssrn.2285042

Adomavicius, G., Mobasher, B., Ricci, F., & Tuzhilin, A. (2011). Context-aware recommender systems. *AI Magazine, 32*(3), 67–80. doi:10.1609/aimag.v32i3.2364

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering, 17*(6), 734–749. doi:10.1109/TKDE.2005.99

Adomavicius, G., & Tuzhilin, A. (2011). Context-aware recommender systems. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender Systems Handbook*, 217–253. Berlin Heidelberg, Germany: Springer.

Aggarwal, P., & Vaidyanathan, R. (2003). Eliciting online customers' preferences: Conjoint vs self-explicated attribute-level measurements. *Journal of Marketing Management, 19*(1–2), 157–177.

Allen, I. E., & Seaman, C. A. (2007). Likert scales and data analyses. *Quality Progress, July*. http://asq.org/quality-progress/2007/07/statistics/likert-scales-and-data-analyses.html.

Amatriain, X., Pujol, J. M., Tintarev, N., & Oliver, N. (2009). Rate it again: increasing recommendation accuracy by user re-rating. In *Proceedings of the third ACM conference on Recommender systems (RecSys '09)*, 173–180. New York, NY, USA: ACM. doi:10.1145/1639714.1639744

Anand, S. S., & Mobasher, B. (2007). Introduction to intelligent techniques for web personalization. *ACM Transactions on Internet Technology, 7*(4), Article No. 18. doi:10.1145/1278366.1278367

Anderson, J., & Rainie, L. (2012). The future of big data. *Pew Research Center's Internet & American Life Project*. http://www.pewinternet.org/~/media//Files/Reports/2012/PIP_Future_of_Internet_2012_Big_Data.pdf

Balasubramanian, S., & Mahajan, V. (2001). The economic leverage of the virtual community. *International Journal of Electronic Commerce, 5*(3), 103–138.

BBC News Online (2001). Sony admits using fake reviewer. http://news.bbc.co.uk/1/hi/entertainment/film/1368666.stm

Berger, J., Sorensen, A. T., & Rasmussen, S. J. (2010). Positive effects of negative publicity: When negative reviews increase sales. *Marketing Science, 29*(5), 815–827. doi:10.1287/mksc.1090.0557

Bernstein, M. S., Marcus, A., Karger, D. R., & Miller, R. C. (2010). Enhancing directed content sharing on the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*, 971–980. New York, NY, USA: ACM. doi:10.1145/1753326.1753470

Bevan, N. (2009). What is the difference between the purpose of usability and user experience evaluation methods? *User Experience Evaluation Methods in Product Development (UXEM'09)* (Workshop in INTERACT 2009). Four pages. http://www.nigelbevan.com/papers/What_is_the_difference_between_usability_and_user_experience_evaluation_methods.pdf.

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems, 46*, 109–132. doi:10.1016/j.knosys.2013.03.012

Bollen, D., Knijnenburg, B. P., Willemsen, M. C., & Graus, M. (2010). Understanding choice overload in recommender systems. In *Proceedings of the fourth ACM conference on Recommender systems (RecSys '10)*, 63–70. New York, NY, USA: ACM. doi:10.1145/1864708.1864724

Brusilovsky, P. (2007). Adaptive navigation support. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The Adaptive Web*, LNCS 4321, 263–290. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-540-72079-9_8

Bryman, A. (2004). Triangulation. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE Encyclopedia of Social Science Research Methods*.

Thousand Oaks, CA, USA: SAGE Publications. http://www.sagepub.com/chambliss4e/study/chapter/encyc_pdfs/4.2_Triangulation.pdf

Buder, J., & Schwind, C. (2012). Learning with personalized recommender systems: A psychological view. *Computers in Human Behavior, 28*(1), 207–216. doi:10.1016/j.chb.2011.09.002

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction, 12*(4), 331–370. doi:10.1023/A:1021240730564

Burke, R. (2007). Hybrid web recommender systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The Adaptive Web*, LNCS 4321, 377–408. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-540-72079-9_12

Bush, V. (1945). As we may think. *The Atlantic*. http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/

Campochiaro, E., Casatta, R., Cremonesi, P., & Turrin, R. (2009). Do metrics make recommender algorithms? In *Proceedings of the 2009 International Conference on Advanced Information Networking and Applications Workshops (WAINA '09)*, 648–653. Washington, DC, USA: IEEE Computer Society. doi:10.1109/WAINA.2009.127

Castagnos, S., Jones, N., & Pu, P. (2009). Recommenders' influence on buyers' decision process. In *Proceedings of the Third ACM Conference on Recommender Systems (RecSys '09)*, 361–364. New York, NY, USA: ACM. doi:10.1145/1639714.1639788

Cena, F., Likavec, S., Lombardi, I., & Picardi, C. (2011). Synthesis of collective tag-based opinions in the socialweb. In R. Pirrone & F. Sorbello (Eds.), *Proceedings of the 12th International Conference on Artificial Intelligence around Man and Beyond (AI*IA'11)*, LNCS 6934, 286–298. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-642-23954-0_27

Cena, F., Vernero, F., & Gena, C. (2010). Towards a customization of rating scales in adaptive systems. In P. Bra, A. Kobsa, & D. Chin (Eds.), *Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization (UMAP'10)*, LNCS 6075, 369–374. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-642-13470-8_34

Chalmers, M., Dieberger, A., Höök, K., & Rudström, Å. (2004). Social navigation and seamful design. *Cognitive Studies, 11*(3), 1–11. http://www.sics.se/~kia/papers/Japan2004.pdf

Chen, L. (2010). Understanding buyers' social information needs during purchase decision process. In *Proceedings of the 12th International Conference on Electronic Commerce: Roadmap for the Future of Electronic Business* (ICEC '10), 15–22. New York, NY, USA: ACM. doi:10.1145/2389376.2389380

Chen, Y.-F. (2008). Herd behavior in purchasing books online. *Computers in Human Behavior, 24*(5), 1977–1992. doi:10.1016/j.chb.2007.08.004

Cheung, C. M. K., & Lee, M. K. O. (2012). What drives consumers to spread electronic word of mouth in online consumer-opinion platforms. *Decision Support Systems, 53*(1), 218–225. doi:10.1016/j.dss.2012.01.015

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research, 43*(3), 345–354. https://msbfile03.usc.edu/digitalmeasures/mayzlin/intellcont/chevalier_mayzlin06-1.pdf

Cleger-Tamayo, S., Fernández-Luna, J. M., & Huete, J. F. (2012). Top-N news recommendations in digital newspapers. *Knowledge-Based Systems,* 27, 180–189. doi:10.1016/j.knosys.2011.11.017

Cone Online Influence Trend Tracker (2011). http://www.slideshare.net/ecommercenews/2011-cone-onlineinfluencetrendtrackerreleaseandfactsheet

Cosley, D., Frankowski, D., Kiesler, S., Terveen, L., & Riedl, J. (2005). How oversight improves member-maintained communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*, 11–20. New York, NY, USA: ACM. doi:10.1145/1054972.1054975

Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., & Riedl, J. (2003). Is seeing believing? How recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*, 585–592. New York, NY, USA: ACM. doi:10.1145/642611.642713

Cremonesi, P., Garzotto, F., Negro, S., Papadopoulos, A. V., & Turrin, R. (2011). Looking for "good" recommendations: a comparative evaluation of recommender systems. In P. Campos, N. Nunes, N. Graham, J. Jorge, & P. Palanque (Eds.), *Proceedings of the 13th IFIP TC 13 international conference on Human-computer interaction*, LNCS 6948, 152–168. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-642-23765-2_11

Cremonesi, P., Garzotto, F., & Turrin, R. (2012a). Investigating the persuasion potential of recommender systems from a quality perspective: An empirical study. *ACM Transactions on Interactive Intelligent Systems, 2*(2), Article No. 11, 41 pages. Doi:10.1145/2209310.2209314

Cremonesi, P., Garzotto, F., & Turrin, R. (2012b). User effort vs. accuracy in rating-based elicitation. In *Proceedings of the sixth ACM conference on Recommender systems (RecSys '12)*, 27–34. New York, NY, USA: ACM. doi:10.1145/2365952.2365963

Cremonesi, P., Garzotto, F., & Turrin, R. (2013). User-centric vs. system-centric evaluation of recommender systems. In P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, & M. Winckler (Eds.), *Proceedings of the 14th IFIP TC 13 International Conference (INTERACT 2013)*, LNCS 8119, 152–168. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-642-23765-2_11

Cremonesi, P., Turrin, R., Lentini, E., & Matteucci, M. (2008). An evaluation methodology for collaborative recommender systems. In *Proceedings of the 2008 International Conference on Automated Solutions for Cross Media Content and Multi-channel Distribution (AXMEDIS '08)*, 224–231. Washington, DC, USA: IEEE Computer Society. doi:10.1109/AXMEDIS.2008.13

Cui, G., Lui, H.-K., & Guo, X. (2012). The effect of online consumer reviews on new product sales. *International Journal of Electronic Commerce, 17*(1), 39–58. doi:10.2753/JEC1086-4415170102

Cyr, D., Hassanein, K., Head, M., & Ivanov, A. (2007). The role of social presence in establishing loyalty in e-Service environments. *Interacting with Computers*, *19*(1), 43–56. doi:10.1016/j.intcom.2006.07.010

Damasio, A. (2006). *Descartes' Error*. London, UK: Vintage.

Das, A. S., Datar, M., Garg, A., & Rajaram, S. (2007). Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*, 271–280. New York, NY, USA: ACM. doi:10.1145/1242572.1242610

Day, P. N., Johnson, G. I., Rohan, C., Riley, C., & Carlisle, M. (2010). User experience at NCR: An organisational overview. In *Proceedings of the 24th BCS Interaction Specialist Group Conference (BCS '10)*, 195–203. Swinton, UK: British Computer Society. http://ewic.bcs.org/upload/pdf/ewic_hci10_paper22.pdf

Diaz, Z. (2014). 21% of Americans who have left reviews, reviewed products without buying or trying them. *YouGov Omnibus Research.* http://today.yougov.com/news/2014/01/22/21-americans-have-reviewed-products-and-services-t/

Dieberger, A. (1997). Supporting social navigation on the World Wide Web. *International Journal of Human-Computer Studies, 46*(6), 805–825. doi:10.1006/ijhc.1996.0111

Dieberger, A., Dourish, P., Höök, K., Resnick, P., & Wexelblat, A. (2000). Social navigation: Techniques for building more usable systems. *ACM Interaction*s, 7(6), 36–45. doi:10.1145/352580.352587

Dinet, J., Chevalier, A., & Tricot, A. (2012). Information search activity: An overview. *European Review of Applied Psychology, 62*(2), 49–62. doi:10.1016/j.erap.2012.03.004

Doerr, C., Blenn, N., Tang, S., & Van Mieghem, P. (2012). Are friends overrated? A study for the social news aggregator Digg.com. *Computer Communications*, *35*(7), 796–809. doi:10.1016/j.comcom.2012.02.001

Dourish, P., & Chalmers, M. (1994). Running out of space: Models of information navigation. *Short paper presented at HCI'94*, two pages. http://www.dcs.gla.ac.uk/~matthew/papers/hci94.pdf

Drachsler, H., Hummel, H. G. K., & Koper, R. (2009). Identifying the goal, user model and conditions of recommender systems for formal and informal learning. *Journal of Digital Information, 10*(2), 4–24.

Drenner, S., Sen, S., & Terveen, L. (2008). Crafting the initial user experience to achieve community goals. In *Proceedings of the 2008 ACM Conference on Recommender Systems* (RecSys '08), 187–194. New York, NY, USA: ACM. doi:10.1145/1454008.1454039

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication, 12*(4), 1143–1168. doi:10.1111/j.1083-6101.2007.00367.x

Erdelez, S., & Rioux, K. (2000). Sharing tools on newspaper web sites: An exploratory study. *Online Information Review, 24*(3), 218–228. doi:10.1108/14684520010341290

Farhi, P. (2013). The Washington Post closes sale to Amazon founder Jeff Bezos. *The Washington Post*. http://articles.washingtonpost.com/2013-10-01/business/42568774_1_jeff-bezos-washington-post-co-katharine-graham

Farzan, R., & Brusilovsky, P. (2006). Social navigation support in a course recommendation system. In V. P. Wade, H. Ashman, & B. Smyth (Eds.), *Proceedings of the 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH'06)*, LNCS 4018, 91–100. Berlin Heidelberg, Germany: Springer. doi:10.1007/11768012_11

Freyne, J., Farzan, R., Brusilovsky, P., Smyth, B., & Coyle, M. (2007). Collecting community wisdom: integrating social search & social navigation. In *Proceedings of the 12th international conference on Intelligent user interfaces (IUI '07)*, 52–61. New York, NY, USA: ACM. doi:10.1145/1216295.1216312

Fu, W.-T., Kannampallil, T., Kang, R., & He, J. (2010). Semantic imitation in social tagging. *ACM Transactions on Computer-Human Interaction, 17*(3), Article No. 12, 37 pages. doi:10.1145/1806923.1806926

Gadanho, S. C., & Lhuillier, N. (2007). Addressing uncertainty in implicit preferences. In *Proceedings of the 2007 ACM Conference on Recommender Systems (RecSys '07)*, 97–104. New York, NY, USA: ACM. doi:10.1145/1297231.1297248

Garland, R. (1991). The mid-point on a rating scale: Is it desirable? *Marketing Bulletin*, *2*, 66–70. http://marketing-bulletin.massey.ac.nz/V2/MB_V2_N3_Garland.pdf

Garrison, D. R., & Kanuka, H. (2004). Blended learning: Uncovering its transformative potential in higher education. *Internet and Higher Education*, 7(2), 95–105. doi:10.1016/j.iheduc.2004.02.001

Gedikli, F., Ge, M., & Jannach, D. (2011). Understanding recommendations by reading the clouds. *Lecture Notes in Business Information Processing, 85*, 196–208. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-642-23014-1_17

Gedikli, F., & Jannach, D. (2010). Rating items by rating tags. In *Proceedings of the 2nd Workshop on Recommender Systems and the Social Web (RSWEB'10)*, 25–32. http://www.researchgate.net/publication/221141099_Workshop_on_recommender_systems_and_the_social_web/file/79e4150a44794e80e4.pdf

Gena, C., Brogi, R., Cena, F., & Vernero, F. (2011). The impact of rating scales on user's rating behavior. In J. A. Konstan, R. Conejo, J. L. Marzo, & N. Oliver (Eds.), *Proceedings of the 19th international conference on User modeling, adaption, and personalization (UMAP'11)*, LNCS 6787, 123–134. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-642-22362-4_11

Ghauth, K. I., & Abdullah, N. A. (2010). Learning materials recommendation using good learners ratings and content-based filtering. *Journal of Educational Technology Research and Development 58*(6), 711–727. doi:10.1007/s11423-010-9155-4

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM, 35*(12), 61–70. doi:10.1145/138859.138867

Golder, S. A., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science, 32*(2), 198–208. doi:10.1177/0165551506062337

Goldstein, N. J., Martin, S. J., & Cialdini, R. B. (2007). *Yes!* London, UK: Profile Books.

Granger, L. (2013). Giving it away for free: advertising, content and the paywall debate. *memeburn*. http://memeburn.com/2013/10/giving-it-away-for-free-advertising-content-and-the-paywall-debate-t4a/

Granić, A., & Adams, R. (2011). User sensitive research in e-learning: exploring the role of individual user characteristics. *Universal Access in the Information Society*, *10*(3), 307–318. doi:10.1007/s10209-010-0207-7

Gretzel, U., & Fesenmaier, D. (2006). Persuasion in recommender systems. *International Journal of Electronic Commerce, 11*(2), 81–100. doi:10.2753/JEC1086-4415110204

Gupta, P., Goel, A., Lin, J., Sharma, A., Wang, D., & Zadeh, R. (2013). WTF: The who to follow service at Twitter. In *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*, 505–514. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. http://www2013.org/proceedings/p505.pdf

Gupte, M., Hajiaghayi, M., Han, L., Iftode, L., Shankar, P., & Ursu, R. M. (2009). News posting by strategic users in a social network. In: S. Leonardi (Ed.), *Workshop on Internet and Network Economics (WINE)*, LNCS 5929, 632–639. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-642-10841-9_65

Hage, H., & Aïmeur, E. (2008). Harnessing learner's collective intelligence: A web 2.0 approach to e-learning. In B. P. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of the 9th international conference on Intelligent Tutoring Systems (ITS '08)*, LNCS 5929, 438–447. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-540-69132-7_47

Halliday, J. (2010). Digg investigates claims of conservative 'censorship'. *theguardian.com*. http://www.theguardian.com/technology/2010/aug/06/digg-investigates-claims-conservative-censorship

Halpin, H., Robu, V., & Shepherd, H. (2007). The complex dynamics of collaborative tagging. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*, 211–220. New York, NY, USA: ACM. doi:10.1145/1242572.1242602

Hangartner, R. (2007). What is the recommender industry? *MobileGroove.com*. http://www.mobilegroove.com/guest-column-what-is-the-recommender-industry-750

Harford, T. (2007). *The Undercover Economist*. London, UK: Abacus.

Harmon, A. (2004). Amazon glitch unmasks war of reviewers. *NY Times: Technology*. http://www.nytimes.com/2004/02/14/technology/14AMAZ.html.

Harper, F. M., Li, X., Chen, Y., & Konstan, J. A. (2005). An economic model of user rating in an online recommender system. In L. Ardissono, P. Brna, & A. Mitrovic (Eds.), *Proceedings of the 10th International Conference on User Modeling (UM'05)*, LNCS 3538, 307–316. Berlin Heidelberg, Germany: Springer. doi:10.1007/11527886_40

Harper, F. M., Li, S. X., Chen, Y., & Konstan, J. A. (2007). Social comparisons to motivate contributions to an online community. In Y. De Kort, W. IJsselsteijn, C. Midden, B. Eggen, & B. J. Fogg (Eds.), *Proceedings of the 2nd international conference on Persuasive technology (PERSUASIVE'07)*, LNCS 4744, 148–159. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-540-77006-0_20

Harrison, S., & Dourish, P. (1996). Re-place-ing space: The roles of place and space in collaborative systems. In M. S. Ackerman (Ed.), *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work (CSCW '96)*, 67–76. New York, NY, USA: ACM. doi:10.1145/240080.240193

Hassenzahl, M., & Tractinsky, N. (2006). User experience – a research agenda. *Behaviour & Information Technology*, *25*(2), 91–97.

Häubl, G., & Trifts, V. (2000). Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Science, 19*(1), 4–21. doi:10.1287/mksc.19.1.4.15178

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet? *Journal of Interactive Marketing, 18*(1), 38–52.

Herlocker, J. L., Konstan, J. A., Borchers, A., & Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. In

*Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, 230–237. New York, NY, USA: ACM. doi:10.1145/312624.312682

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*, 241–250. New York, NY, USA: ACM. doi:10.1145/358916.358995

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems, 22*(1), 5–53. doi:10.1145/963770.963772

Hill, W. C., Hollan, J. D., Wroblewski, D., & McCandless, T. (1992). Edit wear and read wear. In P. Bauersfeld, J. Bennett, & G. Lynch (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '92)*, 3–9. New York, NY, USA: ACM. doi:10.1145/142750.142751

Hill, W., Stead, L., Rosenstein, M., & Furnas, G. (1995). Recommending and evaluating choices in a virtual community of use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95)*, 194–201. New York, NY, USA: ACM Press. doi:10.1145/223904.223929

Hu, N., Liu, L., & Zhang, J. J. (2008). Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. *Information Technology & Management, 9*(3), 201–214. doi:10.1007/s10799-008-0041-2

Hu, N., Pavlou, P. A., & Zhang, J. (2006). Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication. In *Proceedings of the 7th ACM Conference on Electronic Commerce (EC '06)*, 324–330. New York, NY, USA: ACM. doi:10.1145/1134707.1134743

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM, 52*(10), 144–147. doi:10.1145/1562764.1562800

Hu, R., & Pu, P. (2011). Enhancing recommendation diversity with organization interfaces. In *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI '11)*, 347–350. New York, NY, USA: ACM. doi:10.1145/1943403.1943462

ISO 9241-11 (1998). *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) -- Part 11: Guidance on Usability*. Geneva, Switzerland: International Organization for Standardization.

ISO 9241-210 (2010). *Ergonomics of Human-System Interaction -- Part 210: Human-centred Design for Interactive Systems*. Geneva, Switzerland: International Organization for Standardization.

Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., & Stumme, G. (2007). Tag recommendations in folksonomies. In J. N. Kok, J. Koronacki, R. Lopez De Mantaras, S. Matwin, D. Mladenič, & A. Skowron (Eds.), *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2007)*, LNCS 4702, 506–514. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-540-74976-9_52

Jeckmans, A. J. P., Beye, M., Erkin, Z., Hartel, P., Lagendijk, R. L., & Tang, Q. (2013). Privacy in recommender systems. N. Ramzan, R. van Zwol, J.-S. Lee, K. Clüver, & X.-S. Hua (Eds.), *Social Media Retrieval*, 263–281. London, UK: Springer. doi:10.1007/978-1-4471-4555-4_12

Jurca, R., & Faltings, B. (2009). Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research, 34*(1)), 209–253.

Juvina, I., & van Oostendorp, H. (2004). Predicting user preferences: From semantic to pragmatic metrics of Web navigation behavior. In *Proceedings of the Conference on Dutch Directions in HCI (Dutch HCI '04)*, four pages. New York, NY, USA: ACM. doi:10.1145/1005220.1005233

Kadet, A. (2007). Rah-rah ratings online. *SmartMoney*. http://www.smartmoney.com/spend/rip-offs/ rah-rah-ratings-online-20837/

Kahneman, D. (2012). *Thinking, Fast and Slow*. London, England: Penguin Books.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–291.

Kalogeraki, S. (2009). The divergence hypothesis in modernization theory across three European countries: The UK, Sweden and Greece. *Culture Unbound 1*, 161–178. http://www.cultureunbound.ep.liu.se/v1/a10/cu09v1a10.pdf

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, *65*(4), 681–706.

Kee, T. (2008). Majority of online shoppers check at least four reviews before buying. *OnlineMediaDaily*. http://www.mediapost.com/publications/article/76727/majority-of-online-shoppers-check-at-least-four-re.html

Kiliç-Çakmak, E. (2010). Learning strategies and motivational factors predicting information literacy self-efficacy of e-learners. *Australasian Journal of Educational Technology, 26*(2), 192–208. http://www.ascilite.org.au/ajet/ajet26/kilik-cakmak.html

Kluver, D., Nguyen, T. T., Ekstrand, M., Sen, S., & Riedl, J. (2012). How many bits per rating? In *Proceedings of the Sixth ACM Conference on Recommender Systems (RecSys '12)*, 99–106. New York, NY, USA: ACM. doi:10.1145/2365952.2365974

Knijnenburg, B. P. (2012). Conducting user experiments in recommender systems. In *Proceedings of the Sixth ACM Conference on Recommender Systems (RecSys '12)*, 3–4. New York, NY, USA: ACM. doi:10.1145/2365952.2365956

Knijnenburg, B. P., Reijmer, N. J. M., & Willemsen, M. C. (2011). Each to his own: How different users call for different interaction methods in recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems (RecSys '11)*, 141–148. New York, NY, USA: ACM. doi:10.1145/2043932.2043960

Knijnenburg, B. P., & Willemsen, M. C. (2009). Understanding the effect of adaptive preference elicitation methods on user satisfaction of a recommender system. In *Proceedings of the Third ACM Conference on Recommender Systems (RecSys '09)*, 381–384. New York, NY, USA: ACM. doi:10.1145/1639714.1639793

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction, 22*(4–5), 441–504. doi:10.1007/s11257-011-9118-4

Knijnenburg, B. P., Willemsen, M. C., & Kobsa, A. (2011). A pragmatic procedure to support the user-centric evaluation of recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*, 321–324. New York, NY, USA: ACM. doi:10.1145/2043932.2043993

Kobsa, A. (2007). Privacy-enhanced web personalization. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The Adaptive Web*, LNCS 4321, 628–670. Berlin Heidelberg, Germany: Springer.

Komiak, S. Y. X., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly, 30*(4), 941–960.

Konstan, J. A. (2008). Introduction to recommender systems. *SIGMOD 2008 Tutorial*. http://www-users.cs.umn.edu/~konstan/SIGMOD-2008-Tut.pdf

Konstan, J. A. (2012). *Recommender Systems: The Power of Personalization Q&A*. http://learning.acm.org/webinar/recommender_qa.cfm

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). GroupLens: Applying collaborative filtering to Usenet news. *Communications of the ACM, 40*(3), 77–87. doi:10.1145/245108.245126

Konstan, J. A., & Riedl, J. (2012). Recommender systems: From algorithms to user experience. *User Modeling and User-Adapted Interaction, 22*, 101–123. doi:10.1007/s11257-011-9112-x.

Konstas, I., Stathopoulos, V., & Jose, J. M. (2009). On social networks and collaborative recommendation. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*, 195–202. New York, NY, USA: ACM. doi:10.1145/1571941.1571977

Krakovsky, M. (2011). All the news that's fit for you. *Communications of the ACM, 54*(6), 20–21. doi:10.1145/1953122.1953129

Kuflik, T., Wecker, A. J., Cena, F., & Gena, C. (2012). Evaluating rating scales personality. In J. Masthoff, B. Mobasher, M. C. Desmarais, & R. Nkambou (Eds.), *Proceedings of the 20th international conference on User Modeling, Adaptation, and Personalization (UMAP'12)*, 310–315. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-642-31454-4_27

Kumar, N., & Benbasat, I. (2006). Research note: The influence of recommendations and consumer reviews on evaluations of websites. *Information Systems Research, 17*(4), 425–439. doi:10.1287/isre.1060.0107

Lam, S. K. T., Frankowski, D., & Riedl, J. (2006). Do you trust your recommendations? An exploration of security and privacy issues in recommender systems. In G. Müller (Ed.), *Proceedings of the 2006 International Conference on Emerging Trends in Information and Communication Security (ETRICS'06)*, LNCS 3995, 14–29. Berlin Heidelberg, Germany: Springer. doi:10.1007/11766155_2

Lam, S. K., & Riedl, J. (2004). Shilling recommender systems for fun and profit. In *Proceedings of the 13th International Conference on World Wide Web (WWW '04)*, 393–402. New York, NY, USA: ACM. doi:10.1145/988672.988726

Landia, N., & Anand, S. S. (2009). Personalised tag recommendation. In *Proceedings of the ACM RecSys09 Workshop on the Recommender Systems and the Social Web,* four pages. http://ls13-www.cs.uni-dortmund.de/homepage/RSWEB/papers/Landia.pdf

Lee, S. E., Son, D. K., & Han, S. S. K. (2007). Qtag: Tagging as a means of rating, opinion-expressing, sharing and visualizing. In *Proceedings of the 25th Annual ACM International Conference on Design of Communication (SIGDOC '07)*, 189–195. New York, NY, USA: ACM. doi:10.1145/1297144.1297184

Leino, J. (2008). *Searches and Recommendations: Item-Finding in Complex Information Environments*. M.Sc. thesis. University of Tampere. http://www.cs.uta.fi/research/theses/masters/Leino_Juha.pdf

Lerman, K. (2007). Dynamics of collaborative document rating systems. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD '07)*, 46–55. New York, NY, USA: ACM. doi:10.1145/1348549.1348555

Lerman, K., & Ghosh, R. (2010). Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In *Proceedings of 4th International AAAI Conference on Weblogs and Social Media (ICWSM 2010)*, 90–97. Menlo Park, CA, USA: AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1509/1839

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76–80. doi:10.1109/MIC.2003.1167344

Ling, K., Beenen, G., Ludford, P., Wang, X., Chang, K., Li, X., Cosley, D., Frankowski, D., Terveen, L., Rashid, A. M., Resnick, P., & Kraut, R. (2005). Using social psychology to motivate contributions to online communities. *Journal of Computer-Mediated Communication, 10*(4), 212–221. http://jcmc.indiana.edu/vol10/issue4/ling.html

Liu, J., Dolan, P., & Pedersen, E. R. (2010). Personalized news recommendation based on click behavior. In *Proceedings of the 15th International Conference on Intelligent User Interfaces (IUI '10),* 31–40. New York, NY, USA: ACM. doi:10.1145/1719970.1719976

López-Pérez, M. V., Pérez-López, M. C., & Rodríguez-Ariza, L. (2011). Blended learning in higher education: Students' perceptions and their relation to outcomes. *Computers & Education, 56*(3), 818–826. doi:10.1016/j.compedu.2010.10.023

Ludford, P. J. (2007). *Technologies for Managing and Sharing Local Information*. Ph.D. Thesis. University of Minnesota Dept. of Computer Science and Engineering. http://www-users.cs.umn.edu/~ludford/dissertation_final.pdf.

Malone, T. W., Laubacher, R., & Dellarocas, C. N. (2010). Harnessing crowds: Mapping the genome of collective intelligence. *MIT Sloan Research Paper No. 4732-09*. doi:10.2139/ssrn.1381502

Maltz, D., & Ehrlich, K. (1995). Pointing the way: Active collaborative filtering. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95)*, 202–209. New York, NY, USA: ACM Press. doi:10.1145/223904.223930

Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., & Koper, R. (2011). Recommender systems in technology enhanced learning. In P. B. Kantor, F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook*, 387–415. Berlin Heidelberg: Springer.

Marinho, L. B., & Schmidt-Thieme, L. (2007). Collaborative tag recommendations. In *Proceedings of 31st Annual Conference of the Gesellschaft für Klassifikation (GfKl)*, LNAI 4702, 506–514. Berlin Heidelberg, Germany: Springer.

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia (HYPERTEXT '06)*, 31–40. New York, NY, USA: ACM. doi:10.1145/1149941.1149949

McNee, S. M., Lam, S. K., Konstan, J. A., & Riedl, J. (2003). Interfaces for eliciting new user preferences in recommender systems. In *Proceedings of the 9th International Conference on User Modeling (UM 2003)*, LNAI 2702, 178–188. Berlin Heidelberg, Germany: Springer.

McNee, S. M., Riedl, J., & Konstan, J. A. (2006a). Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)*, 1097–1101. New York, NY, USA: ACM. doi:10.1145/1125451.1125659

McNee, S. M., Riedl, J., & Konstan, J. A. (2006b). Making recommendations better: An analytic model for human-recommender interaction. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)*, 1103–1108. New York, NY, USA: ACM. doi:10.1145/1125451.1125660

Mitchell, A., Kiley, J., Gottfried, J., & Guskin, E. (2013). The role of news on facebook: Common yet incidental. *PewResearchCenter*.

http://www.journalism.org/files/2013/10/facebook_news_10-24-2013.pdf

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly, 34*(1), 185–200.

Neumann, A. W. (2007). Motivating and supporting user interaction with recommender systems. In L. Kovács, N. Fuhr, & C. Meghini (Eds.), *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'07)*, LNCS 4675, 428–439. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-540-74851-9_36

Nguyen, A.-T., Denos, N., & Berrut, C. (2007). Improving new user recommendations with rule-based induction on cold user data. In *Proceedings of the 2007 ACM Conference on Recommender Systems (RecSys '07)*, 121–128. New York, NY, USA: ACM. doi:10.1145/1297231.1297251

Nov, O., & Ye, C. (2010). Why do people tag? Motivations for photo tagging. *Communications of the ACM, 53*(7), 128–131. doi:10.1145/1785414.1785450

O'Brien, J. M. (2006). The race to create a 'smart' Google. *FORTUNE Magazine*. http://money.cnn.com/magazines/fortune/fortune_archive/2006/11/27/8394347/

O'Mahony, M. P., Hurley, N. J., & Silvestre, G. C. M. (2006). Detecting noise in recommender system databases. In *Proceedings of the 11th International Conference on Intelligent User Interfaces (IUI '06)*, 109–115. New York, NY, USA: ACM. doi:10.1145/1111449.1111477

O'Toole, J. (2013). Amazon's Bezos buys Washington Post for $250 million. *CNN Money*. http://money.cnn.com/2013/08/05/news/companies/washington-post-bezos/

Ozok, A. A., Fan, Q., & Norcio, A. F. (2010). Design guidelines for effective recommender system interfaces based on a usability criteria conceptual model: results from a college student population. *Behaviour & Information Technology, 29*(1), 57–83. doi:10.1080/01449290903004012

Packer, G. (2014). Cheap words. *The New Yorker*. http://www.newyorker.com/reporting/2014/02/17/140217fa_fact_packer?printable=true&currentPage=all

Park, D. H., Kim, H. K., Choi, I. Y., & Kim, J. K. (2012). A literature review and classification of recommender systems research. *Expert Systems with Applications, 39*(11), 10059–10072. doi:10.1016/j.eswa.2012.02.038

Pathak, B., Garfinkel, R., Gopal, R., Venkatesan, R., & Yin, F. (2010). Empirical analysis of the impact of recommender systems on sales. *Journal of Management Information Systems, 27*(2), 159–188. doi:10.2753/MIS0742-1222270205

Pommeranz, A., Broekens, J., Wiggers, P., Brinkman, W.-P., & Jonker, C. M. (2012). Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process. *User Modeling and User-Adapted Interaction, 22*(4–5), 357–397. doi:10.1007/s11257-011-9116-6

Pommeranz, A., Wiggers, P., Jonker, C. M. (2011). Towards compositional design and evaluation of preference elicitation interfaces. In M. Kurosu (Ed.), *Proceedings of the 2nd International Conference on Human Centered Design (HCD'11)*, LNCS 6776, 586–596. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-642-21753-1_65

Pu, P., Chen, L., & Hu, R. (2011). A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM Conference on Recommender Systems (RecSys '11)*, 157–164. New York, NY, USA: ACM. doi:10.1145/2043932.2043962

Pu, P., Chen, L., & Hu, R. (2012). Evaluating recommender systems from the user's perspective: Survey of the state of the art. *User Modeling and User-Adapted Interaction*, *22*, 317–355. doi:10.1007/s11257-011-9115-7

Purcell, K., Rainie, L., Mitchell, A., Rosenstiel, T., & Olmstead, K. (2010). Understanding the participatory news consumer. *Pew Internet & American Life Project*. http://www.pewinternet.org/2010/03/01/understanding-the-participatory-news-consumer/

Racherla, P., & Friske, W. (2012). Perceived 'usefulness' of online consumer reviews: An exploratory investigation across three services categories. *Electronic Commerce Research and Applications, 11*(6), 548–559. doi:10.1016/j.elerap.2012.06.003

Rashid, A. M., Albert, I., Cosley, D., Lam, S. K., McNee, S. M., Konstan, J. A., & Riedl, J. (2002). Getting to know you: Learning new user preferences in recommender systems. In *Proceedings of the 7th International Conference on Intelligent User Interfaces (IUI '02)*, 127–134. New York, NY, USA: ACM. doi:10.1145/502716.502737

Rashid, A. M., Ling, K., Tassone, R. D., Resnick, P., Kraut, R., & Riedl, J. (2006). Motivating participation by displaying the value of contribution. In R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, & G. Olson (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*, 955–958. New York, NY, USA: ACM. doi:10.1145/1124772.1124915

Recker, M. M., Walker, A., & Lawless, K. (2003). What do you recommend? Implementation and analyses of collaborative information filtering of web resources for education. *Instructional Science, 31*, 299–316.

Redpath, J., Glass, D. H., McClean, S., & Chen, L. (2010). Collaborative filtering: The aim of recommender systems and the significance of user ratings. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, & S. Little (Eds.), *Proceedings of the 32nd European Conference on Advances in Information Retrieval (ECIR'2010)*, 394–406. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-642-12275-0_35

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work (CSCW '94)*, 175–186. New York, NY, USA: ACM. doi:10.1145/192844.192905

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM, 40*(3), 56–58. doi:10.1145/245108.245121

Riedl, J., & Dourish, P. (2005). Introduction to the special section on recommender systems. *ACM Transactions on Computer-Human Interaction, 12*(3), 371–373. doi:10.1145/1096737.1096738

Rokach, L. (2009). Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. *Computational Statistics & Data Analysis, 53*(12), 4046–4072. doi:10.1016/j.csda.2009.07.017

Rubinstein, A. (1998). *Modeling Bounded Rationality*. Cambridge, MA, USA: The MIT Press.

Ruohotie, P., & Nokelainen, P. (2003). Practical considerations of motivation and computer-supported collaborative learning. In T. Varis, T. Utsumi, & W.R. Klemm (Eds.), *Global Peace through the Global University System*, 226–236. Hämeenlinna, Finland: University of Tampere.

Said, A., Jain, B. J., Narr, S., & Plumbaum, T. (2012). Users and noise: The magic barrier of recommender systems. In J. Masthoff, B. Mobasher, M. C. Desmarais, & R. Nkambou (Eds.), *Proceedings of the 20th*

*International Conference on User Modeling, Adaptation, and Personalization (UMAP'12)*, LNCS 7379, 237–248. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-642-31454-4_20

Santos, O. C., & Boticario, J. G. (2008). Users' experience with a recommender system in an open source standard-based learning management system. In A. Holzinger (Ed.), *HCI and Usability for Education and Work,* LNCS 5298, 185–204. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-540-89350-9_14

Santos, O. C., & Boticario, J. G. (2010). Modeling recommendations for the educational domain. *Procedia Computer Science, 1*(2), 2793–2800. doi:10.1016/j.procs.2010.08.004

Schafer, J.B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The Adaptive Web*, LNCS 4321, 291–324. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-540-72079-9_9

Schafer, J. B., Konstan, J., & Riedl, J. (1999). Recommender systems in e-commerce. In *Proceedings of the 1st ACM Conference on Electronic Commerce (EC '99)*, 158–166. New York, NY, USA: ACM. doi:10.1145/336992.337035

Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-Commerce recommendation applications. *Data Mining and Knowledge Discovery, 5*(1–2), 115–153. doi:10.1023/A:1009804230409

Schclar, A., Tsikinovsky, A., Rokach, L., Meisels, A., & Antwarg, L. (2009). Ensemble methods for improving the performance of neighborhood-based collaborative filtering. In *Proceedings of the Third ACM Conference on Recommender Systems (RecSys '09)*, 261–264. New York, NY, USA: ACM. doi:10.1145/1639714.1639763

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93–105. doi:10.1037/0003-066X.54.2.93

Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology, 14*(4), 332–348. doi:10.1207/s15327663jcp1404_2

Schwind, C., Buder, J., & Hesse, F. W. (2011). I will do it, but I don't like it: User reactions to preference-inconsistent recommendations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*, 349–352. New York, NY, USA: ACM. doi:10.1145/1978942.1978992

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M., & Riedl, J. (2006). Tagging, communities, vocabulary, evolution. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work (CSCW '06)*, 181–190. New York, NY, USA: ACM. doi:10.1145/1180875.1180904

Sen, S., Vig, J., & Riedl, J. (2009). Tagommenders: Connecting users to items through tags. In *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*, 671–680. New York, NY, USA: ACM. doi:10.1145/1526709.1526800

Shardanand, U., & Maes, P. (1995). Social information filtering: Algorithms for automating "word of mouth". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95)*, 210–217. New York, NY, USA: ACM Press. doi:10.1145/223904.223931

Sillence, E., & Briggs, P. (2007). Please advise: Using the internet for health and financial advice. *Computers in Human Behavior*, *23*(1), 727–748. doi:10.1016/j.chb.2004.11.006

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. http://www.math.mcgill.ca/vetta/CS764.dir/bounded.pdf

Simon, H. A. (1996). *The Sciences of the Artificial* (3rd ed.). Cambridge, MA, US: The MIT Press.

Sinha, R., & Swearingen, K. (2002). The role of transparency in recommender systems. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems (CHI EA '02)*, 830–831. New York, NY, USA: ACM. doi:10.1145/506443.506619

Solsman, J. E. (2013). Bezos to Post: You too can be like Amazon, in 3 simple steps. *CNET*. http://news.cnet.com/8301-1023_3-57601060-93/bezos-to-post-you-too-can-be-like-amazon-in-3-simple-steps/

Sparling, E. I., & Sen, S. (2011). Rating: How difficult is it? In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*, 149–156. New York, NY, USA: ACM. doi:10.1145/2043932.2043961

Stefani, A., Vassiliadis, B., & Xenos, M. (2006). Behavioral patterns in hypermedia systems: A short study of e-commerce vs. e-learning practices. In: *Adaptive and Personalized Semantic Web Studies in Computational Intelligence, 14*, 57–64. Berlin Heidelberg, Germany: Springer. doi:10.1007/3-540-33279-0_4

Streitfeld, D. (2012). The best book reviews money can buy. *NY Times: Business Day*. http://www.nytimes.com/2012/08/26/business/book-reviewers-for-hire-meet-a-demand-for-online-raves.html/s

Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence, 4*, 19 pages. doi:10.1155/2009/421425

Sun, S. (2012). How does the variance of product ratings matter? *Management Science, 58*(4), 696–707. doi:10.1287/mnsc.1110.1458

Svensson, M., Höök, K., & Cöster, R. (2005). Designing and evaluating Kalas: A social navigation system for food recipes. *ACM Transactions on Computer-Human Interaction, 12*(3), 374–400. doi:10.1145/1096737.1096739

Swearingen, K., & Sinha, R. (2001). Beyond algorithms: An HCI perspective on recommender systems. In *Proceedings of the SIGIR 2001 Workshop on Recommender Systems*. New York, NY, USA: ACM. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=0?doi=10.1.1.3.8339&rep=rep1&type=pdf

Swearingen, K., & Sinha, R. (2002). Interaction design for recommender systems. In *Designing Interactive Systems (DIS2002)*. New York, NY, USA: ACM. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=0?doi=10.1.1.15.7347&rep=rep1&type=pdf

Talwar, A., Jurca, R., & Faltings, B. (2007). Understanding user behavior in online feedback reporting. In *Proceedings of the 8th ACM Conference on Electronic Commerce (EC '07)*, 134–142. New York, NY, USA: ACM. doi:10.1145/1250910.1250931

Tang, T., & McCalla, G. (2004a). Beyond learners' interest: Personalized paper recommendation based on their pedagogical features for an e-learning system. In C. Zhang, H. W. Guesgen, & W.-K. Yeap (Eds.), *PRICAI 2004: Trends in Artificial Intelligence*, LNCS 3157, 301–310. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-540-28633-2_33

Tang, T., & McCalla, G. (2004b). Utilizing artificial learners to help overcome the cold-start problem in a pedagogically oriented paper recommendation system. In C. Zhang, H. W. Guesgen, & W.-K. Yeap (Eds.), *PRICAI 2004: Trends in Artificial Intelligence*, LNCS 3137, 245–254. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-540-27780-4_28

Tang, T., & McCalla, G. (2009). A multidimensional paper recommender: Experiments and evaluations. *IEEE Internet Computing, 13*(4), 34–41. doi:10.1109/MIC.2009.73

Terveen, L., & Hill, W. (2001). Beyond recommender systems: Helping people help each other. In J. Carroll (Ed.), *HCI in the New Millennium,* 487–509. Boston, MA, USA: Addison-Wesley.

Thatcher, A. (2006). Information-seeking behaviours and cognitive search strategies in different search tasks on the WWW. *International Journal of Industrial Ergonomics, 36*(12), 1055–1068. doi:10.1016/j.ergon.2006.09.012

Thor, A., Golovin, N., & Rahm, E. (2005). Adaptive website recommendations with AWESOME. *The VLDB Journal, 14*(4), 357–372. doi:10.1007/s00778-005-0160-x

Tso-Sutter, K. H. L., Marinho, L. B., & Schmidt-Thieme, L. (2008). Tag-aware recommender systems by fusion of collaborative filtering algorithms. In *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC '08)*, 1995–1999. New York, NY, USA: ACM. doi:10.1145/1363686.1364171

Vermeir, I., Van Kenhove, P., & Hendrickx, H. (2002). The influence of need for closure on consumer's choice behaviour. *Journal of Economic Psychology, 23*, 703–727. doi:10.1016/S0167-4870(02)00135-6

Vig, J., Sen, S., & Riedl, J. (2009). Tagsplanations: Explaining recommendations using tags. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI '09)*, 47–56. New York, NY, USA: ACM. doi:10.1145/1502650.1502661

Wal, T. V. (2007). Tagging that works. http://s3.amazonaws.com/2007presentations/web20expo/Tagging_that_works_final.pdf

Wei, C.-P., Shaw, M. J., & Easley, R. F. (2002). A survey of recommendation systems in electronic commerce. In Rust, R. T., & Kannan, P. K. (Eds.), *E-Service: New Directions in Theory and Practice*. Armonk, NY, USA: M. E. Sharpe Publisher.

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement, 34*(2), 105–121. doi:10.1177/0146621609338593

Williams, E. (2005). Ten rules for web startups. *evhead*. http://evhead.com/2005/11/ten-rules-for-web-startups.asp

Xiao, B., & Benbasat, I. (2007). E-commerce product recommendation agents: Use, characteristics, and impact. *MIS Quarterly, 31*(1), 137–209. http://www.misq.org/e-commerce-product-recommendation-agents-use-characteristics-and-impact.html

Yang, J., & Mai, E. (2010). Experiential goods with network externalities effects: An empirical study of online rating system. *Journal of Business Research, 63*, 1050–1057. doi:10.1016/j.jbusres.2009.04.029

Zhang, J. (2011). Anchoring effects of recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*, 375–378. New York, NY, USA: ACM. doi:10.1145/2043932.2044010

Ziegler, C.-N., McNee, S. M., Konstan, J. A., & Lausen, G. (2005). Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web (WWW '05)*, 22–32. New York, NY, USA: ACM. doi:10.1145/1060745.1060754

# Publication I

Leino, J., & Räihä, K.-J. (2007). Case Amazon: Ratings and reviews as part of recommendations. In *Proceedings of the 2007 ACM Conference on Recommender Systems (RecSys '07)*, 137–140. New York, NY, USA: ACM. doi:10.1145/1297231.1297255

# Publication II

Leino, J., & Räihä, K.-J. (2008). User experiences and impressions of recommenders in complex information environments. *IEEE Data Engineering Bulletin 31*(2), 32–39. http://sites.computer.org/debull/A08June/raiha.pdf

# Publication III

Leino, J., Räihä, K.-J., & Finnberg, S. (2011). All the news that's fit to read: Finding and recommending news online. In P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, & M. Winckler (Eds.), *13th IFIP TC 13 International Conference (INTERACT 2011)*, LNCS 6948, 169–186. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-642-23765-2_12

# Publication IV

Leino, J. (2012). Case study: Material additions, ratings, and comments in a course setting. In O. C. Santos & J. G. Boticario (Eds.), *Educational Recommender Systems and Technologies: Practices and Challenges*, 258–280. Hershey, PA, USA: IGI Global. doi:10.4018/978-1-61350-489-5.ch011

# Publication V

Leino, J. (2012) Case study: Recommending course reading materials in a small virtual learning community. *International Journal of Web Based Communities 8*(3), 285–301. doi:10.1504/IJWBC.2012.048053

# Publication VI

Leino, J. (2013) Recommending additional study materials: Binary ratings vis-à-vis five-star ratings. In *Proceedings of the 27th International British Computer Society Human Computer Interaction Conference (HCI'13, London, UK)*, Article ID 23101, 10 pages. Swinton, UK: British Computer Society.

# Publication VII

Leino, J., & Heimonen, T. (2013). Improving evaluation honesty and user experience in e-learning by increasing evaluation cost and social presence. In P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, & M. Winckler (Eds.*), 14th IFIP TC 13 International Conference (INTERACT 2013)*, LNCS 8118, 597–615. Berlin Heidelberg, Germany: Springer. doi:10.1007/978-3-642-40480-1_42

**Publications in Dissertations in Interactive Technology series**

Details of the dissertations are available at
http://www.uta.fi/sis/tauchi/dissertations.html.

1. **Timo Partala:** Affective Information in Human-Computer Interaction
2. **Mika Käki:** Enhancing Web Search Result Access with Automatic Categorization
3. **Anne Aula:** Studying User Strategies and Characteristics for Developing Web Search Interfaces
4. **Aulikki Hyrskykari:** Eyes in Attentive Interfaces: Experiences from Creating iDict, a Gaze-Aware Reading Aid
5. **Johanna Höysniemi:** Design and Evaluation of Physically Interactive Games
6. **Jaakko Hakulinen:** Software Tutoring in Speech User Interfaces
7. **Harri Siirtola:** Interactive Visualization of Multidimensional Data
8. **Erno Mäkinen:** Face Analysis Techniques for Human-Computer Interaction
9. **Oleg Špakov:** iComponent – Device-Independent Platform for Analyzing Eye Movement Data and Developing Eye-Based Applications
10. **Yulia Gizatdinova:** Automatic Detection of Face and Facial Features from Images of Neutral and Expressive Faces
11. **Päivi Majaranta:** Text Entry by Eye Gaze
12. **Ying Liu:** Chinese Text Entry with Mobile Phones
13. **Toni Vanhala:** Towards Computer-Assisted Regulation of Emotions
14. **Tomi Heimonen:** Design and Evaluation of User Interfaces for Mobile Web Search
15. **Mirja Ilves:** Human Responses to Machine-Generated Speech with Emotional Content
16. **Outi Tuisku:** Face Interface
17. **Juha Leino:** User Factors in Recommender Systems: Case Studies in e-Commerce, News Recommending, and e-Learning