User Technology Adoption Issues in Recommender Systems

Nicolas Jones
Human Computer Interaction Group
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland
+41 21 6931203
nicolas.jones@epfl.ch

Pearl Pu
Human Computer Interaction Group
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland
+41 21 6936081
pearl.pu@epfl.ch

Abstract

Two music recommender websites, *Pandora* (a content-based recommender) and *Last.fm* (a rating-based social recommender), were compared side-by-side in a within-subject user study involving 64 participants. The main objective was to investigate users initial adoption of recommender technology and their subjective perception of the respective systems. Results show that a simple interface design, the requirement of less initial effort, and the quality of recommended items (accuracy, novelty and enjoyability) are some of the key design features that such websites rely on to break the initial entrance barrier in becoming a popular website.

1 Introduction

As personalized e-commerce websites are gaining popularity, various recommender technologies play an increasingly important role in powering such websites to help users find and discover items that they would like to purchase. However, what makes one website attract millions of users in just few short years while others fail is currently more an art than a science. In general, positive user experiences in terms of user benefits and an easy-to-use site design are associated with a website's popularity and consumer loyalty. Maximizing user benefits entails offering a wide range of system features covering both recommendation and other related services. Although this seems the most logical way to attract users, this approach can directly conflict with the simplicity requirement that underlies the ease-of-use of a website. Therefore characterizing the benefits that users truly want from recommender systems and determining easy-to-use design features, and ultimately finding the balance between these two opposing factors are highly relevant to understanding the usability of recommender systems (RS) used in e-commerce environments.

To explore some of these questions, we began by dividing user experience issues into two broad areas: what motivates users to initially join a recommender website (adoption) and later what motivates them to stay and keep receiving recommendations (loyalty). We started the investigation of the first issue by designing an extensive user study in early 2006, aiming at revealing some of the factors influencing recommender systems' ability to attract new users. We conducted the experiment for 4 weeks, followed by three months of work to organize data, identify the right statistical methods with which to analyze the data and finally derive sound conclusions from the results.

Our user study is the first that compared recommender systems based on two different technologies. Previously researchers have compared only rating based recommender systems [19]. As for the evaluation techniques, we used the same within-subject design approach as found in [18]. This design has the advantage that negative effects such as users' biases and their own propensity for preferring music recommendations and technologies can be maximally reduced. In order to eliminate the influence of learning and fatigue as a result of repeated and extended evaluation, we alternated the order of the two systems for each evaluation and we also used a RM-ANOVA test to confirm that users did not influence each other in terms of their opinions between the groups. As in a typical comparative user study, we first identified the independent variables, which were 1) the

source of recommendations: from *Pandora* or *Last.fm*; and 2) the system itself. To determine the dependent variables, we first focused on those aspects that new users are most likely to experience and classifed them into six particular areas: 1) the initial effort for users to specify their preferences, 2) satisfaction of the interface, 3) enjoyability of the recommended songs, 4) discovery of new songs, 5) perceived accuracy of the RS relative to recommendation provided by friends, and 6) preference of the two systems. In order to verify if users' subjective experiences corresponded to their actual experiences, we also decided to measure several objective variables such as the number of songs they loved, would like to purchase, and disliked. Finally to factor out the elements that most influence users' final preferences (do they prefer *Pandora* or *Last.fm*), we performed correlation analyses (see section 6.3).

In choosing the systems to evaluate, we were influenced by a, at that time, recent blog publication.[1] It compared the features of *Pandora*[2] and *Last.fm*[3], two music recommender systems that employ rather different technologies: *Pandora* is a content-based recommender and *Last.fm*, on the other hand, is a rating-based collaborative filtering recommender. We adopted the use of these two systems for our experiment because they serve our purpose of comparing recommender systems and evaluating their ability to attract new users, while employing two different technologies. We are in no way affiliated with either of the companies providing these systems. Both were contacted without success, in view of establishing a collaboration.

The contribution of this paper lies in first steps for understanding user experience issues with recommender systems, and as the first-stage work the understanding of users' attitudes towards the initial adoption of technology. The final outcome of the work is a broad range of observations and an attempt to produce a set of design guidelines for building effective RS which help achieving the aim of attracting new users. This paper is a first and vital analysis in comparing two recommender systems which use different technologies in the field of recommendations, and aims to evaluate these systems as a whole and not solely reduce them to their background algorithmic nature. The paper above all tries to open a path into this vast problematic and tries to highlight some general principles. However, it must be said that this work has no pretention of affirming that these first highlighted dimensions are *the* key user issues involved in usability and adoption of recommender systems.

The rest of the paper is organized as follows. We first analyze the two contexts in which users may seek or obtain recommendations. This difference in context is critical in helping us understand the right usability issues in the appropriate context. We then review the state-of-the-art of the different recommendation technologies and compare this work with related works that examine system-user interaction issues in recommender systems. We describe *Pandora* and *Last.fm*, the two systems being compared in our user study. We discuss the user study setting, the main results of the experiment, and correlation analysis of measured variables as well as some users' comments. We provide our conclusion followed by our anticipated future work.

2 Context of Recommendation: Seeking vs. Giving

The understanding of usability issues of RS begins with an analysis of why users come to such systems for recommendations. Historically, recommendation technology was only used in recommendation-giving sites where the system observes a user's behavior and learns about his interests and tastes in the background. The system then proposes items that may interest a potential buyer based on the observed history. Therefore users were *given* recommendations as a result of items that they had rated or bought (purchase was used as an indication of preference). In this regard, recommendations are offered as a value-added service to increase the site's ability to attract new users and more importantly obtain their loyalty. At present, an increasingly large number of users go to websites to *seek* advice and suggestions for electronic products, vacation destinations, music, books, etc. They interact with such systems as first-time customers, without necessarily having established a history. Therefore, if we use such technologies to empower recommendation-seeking websites to attract new users, we will encounter user experience problems. Even though alternative methods exist for adapting such technologies for new users, our evaluation of *Last.fm* suggests that users are not inclined to expend effort in establishing history and this initial effort requirement affects their subjective attitudes toward adopting recommendation technologies and their subsequent behaviors towards such systems.

[1] See Krause: Pandora and Last.fm: Nature vs. Nurture in Music Recommenders. http://www.stevekrause.org/

[2] www.pandora.com

[3] www.last.fm

3 Background and Related Work

There has been a great deal of literature produced about recommendation technologies and the comparison of them, especially regarding their technical performances. We briefly describe two technologies about the systems we are evaluating, content and collaborative filtering based technologies. For a more detailed description and comparison of the various technologies in this field, please refer to [1, 5, 16].

3.1 Content-based Recommendation

Content-based recommendation technology has its roots in information retrieval and information filtering. Each item in a database is characterized by a set of attributes, known as the content profile. Such a profile is used to determine if the item is "similar" to the item that a user has preferred in the past and therefore its appropriateness for recommendation. The content profile is constructed by extracting a set of features from an item. In domains such as text documents and electronic products, keywords of a document or physical features of a product are used to build such item profiles and often no further extraction is needed. There are two kinds of context-based recommender systems. In the rating-based approach, each user has an additional user profile which is constructed based on items that he has preferred in the past. Such items are then correlated with other users who have preferred similar items. Therefore such approaches also imply the use of filtering techniques in order to classify the items into groups. Then recommendation is made based on the similitude of groups of items. This is sometimes called the item-to-item recommendation technology.

In another content-based approach, an item liked by a user is taken as his preferred model. This reference item together with the preference model is then used to retrieve "similar" items. In current literature, such systems are known as knowledge-based and conversational recommenders [5] and preference-based product search with example critiquing interfaces [13, 21]. Recommendations are constructed based on users' explicitly stated preferences as they react to a set of examples shown to them. In one variation, the system works as follows: it first deduces a user's preferences by asking him to show the system an example of what she likes. A set of features are then derived from this example to establish a preference model which is then used to generate one or a list of candidates that may interest the user. After viewing the candidates, the user either picks an item or wants to further improve the recommendation quality by critiquing the examples she liked or disliked. The simplest forms of critiques are item-based. More advanced critiquing systems also allow users to build and combine critiques on one or several features of a given item (see critique unit and modality in [6]). This type of recommendation systems does aim to establish long-term generalizations about their users. Instead, it retrieves items that match users' explicitly stated preferences and their critiques.

One requirement of this type of recommender system is that all items must first be encoded into a set of features called the item profile. In most electronic catalogs used in e-commerce environments, products are encoded by the physical features such as the processor speed, the screen size, etc. in the case of portable PCs. This has significantly alleviated the time-consuming task of encoding the item profiles. Due to the difficulty of such tasks, the general belief is that example critiquing-based recommender systems are not feasible for domains such as music where items are not easily amenable to meaningful feature extraction. Another shortcoming of such systems is over-specialization. Users tend to be provided with recommendations that are restricted to what they have specified in their preference models. However, several researchers have developed techniques to overcome this limitation by considering diversity [9, 22] or proposing attractive items that users did not specify (called suggestion techniques) [14].

3.2 Collaborative Filtering Technology

Rather than computing the similarity of items, the collaborative filtering techniques compute correlations among similar users, or "nearest neighbors". Prediction of the attractiveness of an unseen item for a given user is computed based on a combination of the rating scores derived from the nearest neighbors. So such systems recommend items from "like-minded" people rather than users' explicitly stated preferences. Originally, collaborative filtering technology was developed to function in the background of an information provider. While observing what users liked or disliked, the system recommends items that may interest a customer. For example, at Amazon.com, the "people who bought this book also bought" was one(s of the earliest commercial adoptions of this technique. Collaborative filtering systems emphasize the automatic way that users' preferences and tastes are acquired while they perform other tasks (e.g. selecting a book), and the persistent way to provide recommendations based not only on a user's current session history (ephemeral) but also his previous sessions [17].

Despite much effort to improve the accuracy of collaborative filtering meth-

ods [4, 12], several problems still remain unsolved in this domain. When new users come to a recommendation website, the system is unlikely to recommend interesting items because it knows nothing about them. This is called the new-user problem. If a system proposes a random set of items to rate, the quality of recommendation still cannot be guaranteed. [15] for example, proposes several methods to carefully select items that may increase the effective of recommendation for new users. However [10] found that the recommendation quality is optimal when users can rate items out of their own selection rather than rating system-proposed items. Another problem is known as the cold start problem. When a new item becomes available in a database which remains unrated, it tends to stay "invisible" to users. Lastly, this recommendation technology gives users little autonomy in making choices. If a user deviates from the interests and tastes of his "group", she has little chance to see items that she may actually prefer. Related to the autonomy issue is the acceptance issue. When recommendations based on a group of users are suggested to a user, she may not be prepared to accept them due to a low level of system transparency. Herlocker et al. have investigated visualization techniques that explain the neighbor ratings and help users to better accept the results [8]. More recently, Bonhard et al. [3] showed ways to improve collaborative filtering based recommender systems by including information on the profile similarity and rating overlap of a given user.

3.3 Hybrid Recommender Systems

The two approaches have their respective strengths and weaknesses. There have been numerous systems developed to take the hybrid approach which uses the strength of one approach to overcome the limitation of the other. [2] described the Fab digital library project at the Stanford University. It is a content-based collaborative recommender that maintains user profiles based on content analysis, but uses these profiles to determine similar users for collaborative recommendation. [7] also described a hybrid recommendation framework where results from information filtering agents based on content analysis can be combined with the opinions of a community of users to produce better recommendations. See [1] for further details on a survey of recommender technologies and possible scenarios for combing these methods. However, the hybrid approach has not been compared to a pure rating-based collaborative approach in a user study.

3.4 Taxonomy of Recommender Systems in E-Commerce

Schafer et al. [17] examined six e-commerce websites employing one or more variations of recommendation technologies to increase the website's revenue. A classification of technologies was proposed along two main criteria: the degree of automation and the degree of persistence. The former refers to the amount of user effort required to generate the recommendations. The level of persistence measures whether the recommendations are generated based on a user's current session only (ephemeral) or on the user's current session together with his history (persistent). Even though the main analyses are still sound today, the research did not address design issues from a users' motivation-to-join perspective. It did not single out the context of use (recommendation giving vs. seeking) as a critical dimension to characterize recommendation technologies and it did not compare content- vs. rating-based technologies to closely examine the new-user problem.

3.5 Interaction Design for Recommender Systems

Swearingen and Sinha [20] examined system-user interaction issues in recommender systems in terms of the types of user input required, information displayed with recommendations, and the system's user interface design qualities such as layout, navigation, color, graphics, and user instructions. Six collaborative filtering based RS were compared in a user study involving 19 users in order to determine the factors that relate to the effective design of recommender systems beyond the algorithms' level. At the same time, the performance of these six online recommendation systems was compared with that of recommendations from friends of the study participants.

The main results were that an effective recommender system inspires trust in a system which has a transparent system logic, points users to new and not-yet-experienced items, and provides details about recommended items and ways to refine recommendations by including or excluding particular genres. Moreover, they indicated that navigation and layout seemed to be strongly correlated with the ease of use and perceived usefulness of a system.

While our focus is also on system-user interaction issues, our experiment design and results differ from theirs in several significant ways. Our results are complimentary but we focused on design simplicity, users' initial effort requirement and the time it takes for them to receive quality recommendations. For this reason, we have chosen to compare two very different systems on these aspects,

Figure 1: A snapshot of Pandora's main GUI with the embedded flash music player.

while they stayed with rating-based recommenders. With a sample size that is three times as large, 64 vs. 19, our results show that new users prefer RS which require less initial effort, provide higher quality recommendations (enjoyability, accuracy, and novelty) and have an interface that is easy and comfortable to use.

4 The Two Music Systems

4.1 Pandora.com

When a new user first visits *Pandora* (figure 1), a flash-based radio station is launched within 10-20 seconds. Without any requirement on registration, you can enter the name of an artist or a song that you like, and the radio station starts playing an audio stream of songs. For each song played, you can give thumbs up or down to refine what the system is recommending to you next. You can start as many stations as you like with a seed that is either the name of an artist or a song. One can sign in immediately, but the system will automatically prompt all new users to sign in after fifteen minutes, whilst continuing to provide music. As a recognized user, the system remembers your stations and is able to recommend more personalized music to you in subsequent visits. From interacting with *Pandora* and in accordance with indications on its website, it appears that this is an example critiquing-based recommender, based on users' explicitly stated preferences. Furthermore, *Pandora* employs hundreds of professional musicians to encode each song in their database into a vector of hundred features. The system is powered by the Music Genome Project, a wide-ranging analysis of music started in 2000 by a group of musicians and music-loving technologists. The concept is to try and encapsulate the essence of music through hundreds of musical attributes (hence the analogy with *genes*). The focus is on properties of each individual song such as harmony, instrumentation or rhythm, and not so much about a genre to which an artist presumably belongs. The system currently includes songs from more than 10'000 artists and has created more than 13 million stations.

It is conceivable that *Pandora* uses both content- and rating-based approaches. However, in the initial phase of using *Pandora* , the system clearly operates in the content based mode, the approach traditionally used for recommending documents and products.

4.2 Last.fm

Last.fm is a music recommender engine based on a massive collection of music profiles. Each music profile belongs to one person and describes his taste in music. *Last.fm* uses these music profiles to make personalized recommendations by matching users with people who like similar music, and generate personalized radio stations (called recommendation radios) for each person. While it is hard to know the exact technology that powers *Last.fm*, we believe that it uses user-to-user collaborative filtering technology from the ways *Last.fm* behaves and based on information on the website. Their slogans further support our belief. However, it is possible that it also relies on some content-based technology in parts. It is a social recommender and knows little about songs' inherent qualities. It functions purely based on users' rating of items (see previous section for a detailed review of this technology).

With *Last.fm*, a user interacts with it by first downloading and installing a small application, i.e. the music player. *Last.fm* also provides a plugin for recording your music profile through a classic music player like iTunes, but can't take

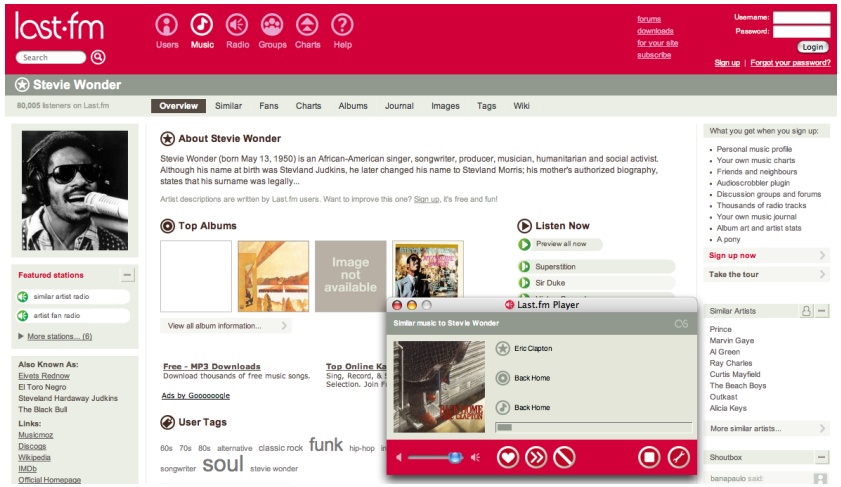

Figure 2: A snapshot of Last.fm's main GUI, with the music player application in foreground.

feedback into account. After the download, you then need to create a user profile which you have to indicate to the player. You can then specify an artist's name, such as "Miles Davis". A list of artists that *Last.fm* believes to be from the same group as Miles Davis will then appear. Now you can listen to an audio stream of songs that belong to that group, and for each song, press a "I like" or "I don't like" button. It is also possible to specify a tag or a set of tags, such as "Indie pop", in the player's interface in order to listen to another suggested stream of audios. Additional features are proposed on the website, as shown on figure 2.

Information from the *Last.fm* website indicates that after a few (∼5) days, a user gets a personalised recommendation radio based on his music profile. According to most of our participants, the songs become relatively interesting and closer to what they like after several hours of listening to the radio stations created in *Last.fm*. In the beginning, the recommendations were not very relevant to their input.

5 Experiment

The experiment was conducted as a within-subject comparative user study with 64 participants (12 females). These were mainly computer and communication science students in their third year at university. There was no financial incentive, but course credit was offered to ensure that the participants were serious about the experiment. Having only computer science students in the study certainly introduced a bias into the data, but was an voluntary choice as usability problems met by such qualified users can only be worse with less skilled computer users.

5.1 Participants' Profiles

The participants' background was made-up of 62% from the 18-24 age group, 34% 25-30 and 4% were above. Subjects' preferred pass-time seems to be "sport" for 32% of the cases, "reading" for 15% and "music" for 47% of them. Since both music RS allow users to buy music, students were initially questioned about their experience with buying songs through internet. Surprisingly, 88% of them had never bought a song or album through an online shop and 6% had only ever bought one song online. Another 3% purchased music sometimes and only the last 3% were regularly customers. A few other questions aimed at determining users' affinity for music were asked, revealing that 44% of the students play an instrument and 30% of those consider themselves musicians (i.e. 13% of all subjects).

5.2 Setup and Procedure

Precise instructions were given throughout the user study and are summarized as follows. In order to complete the experiment, students followed steps provided by a website. The main task was to setup and listen to one radio system for one hour and then answer an online questionnaire of thirty-one questions. Background information was obtained through an initial questionnaire of nine questions. One week later, students tested the other system with the same main questionnaire. At the end, seven preference questions were asked. These steps are detailed hereafter.

The same time limit was given to evaluations to ensure that results would be comparable. In order to maximize the chances that each student would not try to directly compare the two tested systems, *Pandora* and *Last.fm*, they were not informed that they would be testing both and were randomly assigned one

system to test as their first assignment. They were further instructed not to share evaluation results with others. Subjects were first briefed on what they would be doing (system names were kept hidden) before being directed to a website where the detailed instructions were given and a system to test was automatically selected. The website was designed to accompany the students through the whole experiment, step by step.

Step1: In order to make sure that students understood the goals of the experiment, they were provided with an introductive text of an estimated reading time of three minutes. It presented them with a summary of the tasks to complete and informed them of the technical requirements. The opportunity was also taken to remind the subjects that they should behave normally, with the intention of reducing outlier results.

Step 2: The users were then asked to answer the initial questions about their background. General information such as gender, age group and familiarity with internet & computers were asked before targeting two aspects: the subjects' inherent attitudes towards music and recommendations.

Step 3: Once they had completed this questionnaire, they were presented with detailed instructions on the system they were about to test. The page provided the list of tasks they should accomplish, a special one-page document, timing indications and a checklist. The special document was created for both recommender systems and gave the subjects a short summary of the system they were about to test (summary based on the official texts provided online by *Pandora* and *Last.fm*), precise details on how to get the application running and a short explanation on how to give feedback with the system. Since both systems functioned in different ways, we defined a common base of functionalities in which we would be interested, and decided to give detailed information in order to maximize possibilities of comparing the two music recommenders.

Step 4: Subjects were then expected to execute the necessary actions to get their designated system up and running, according to the instructions, before being left to listen to the suggested music during the remaining time. Finally, at the end of the hour, the website automatically redirected the students to the main online questionnaire.

One week later, participants were asked to evaluate the system they had not yet tested. Subjects were not asked to re-state their background, but otherwise the testing procedure was precisely the same. Once finished, seven *preference* questions were added at the end of the main questionnaire. These were aimed at summarizing the subjects' opinions and giving us their preferences on seven selected aspects of the tested radios.

Before starting to listen to music, participants were handed a paper template, designed to help them log and analyze their experience with the system. The main part of the template also allowed users to log each song with its title and the name of the artist. Above all, they could indicate if a song was *new* to them, if they *liked it* or *hated it*, and if they would be prepared to *buy it online* given the opportunity (hereafter new, love, hate and buy).

5.2.1 Questionnaires

The entire user experiment was divided into three parts: the profile questionnaire, the main questionnaire, and the questionnaire that assessed users' preferences of the two systems being compared, which was completed once both systems had been tested.

The main questionnaire questions were chosen according to a set of criteria and hypotheses. In order to evaluate the two music recommender systems, four main themes were defined:

- interface quality
- objective variables
- subjective variables
- user preferences in systems

These domains of questions were chosen in order to provide answers to the main issues of this user study: the effectiveness of RS in terms of its interface quality, the quality of recommended items (accuracy, novelty, enjoyability) relative to its requirements on the users (time to register and download software, time to recommendation) and users' attitudes in adopting the underlying recommender technologies. Additionally, initial effort was investigated through a small set of mixed questions.

6 Results and analysis

6.1 Participants' Background

Besides the demographic background information already reported, results show that the participants possess a high degree of preference for music. This leads us to believe that our subjects are particularly discerning in their assessment of the tested systems thanks to their strong interest in music.

When asked if they had any confidence in computers accurately predicting songs they would like, the subjects were surprisingly positive. 40% of subjects answered "maybe", 35% "cautiously yes" and 12% were "definitely" convinced that computers would be able to recommend songs with precision, leaving only 13% of users unconvinced. Before the study, only one person had heard of *Pandora.com*, and none of *Last.fm*. This assures that the results do not carry much prior bias towards these two systems.

6.2 Main Results

6.2.1 Initial effort

A subset of questions were at first designed to measure users' task time in setting up the respective recommender systems (download and registration time) and the time it takes for a user to receive useful recommendations (time to recommendation). In the allocated one hour of time intented for evaluating the systems, subjects were asked to mark this initial setup time. However, due to the significant difference between the setup cost required by *Pandora* and *Last.fm*, most users were confused and did not record the data as requested. Therefore, we had to resort to facts and user interviews to analyze the initial effort. For *Pandora*, the time to get the flash plug-in and to register is around 2-5 minutes, although you may start listening to music without immediately registering. As for *Last.fm*, the time to download, install the audio player application and to register is 5-15 minutes. However, to get a personalized recommendation radio, a new user has to build up his profile and wait for an average of five days after the registration for his profile to be updated. To conclude, the initial effort required by *Pandora* is only a few minutes, whereas *Last.fm* requires more than few days for users to get started.

It is clear that this *update time* of a few days means that the users of this study didn't have the complete opportunities to enjoy *Last.fm*'s fully personalised radio recommendations. However, this limitation was voluntarily kept as it reflects the computational complexity of the underlying algorithm. Furthermore the experiment intended to evaluate adoption mechanisms and using a bypass to this aspect would have reduced the experiment's interest.

6.2.2 Interface quality

As explained in the experiment setup, several questions were asked on users' experience with the interfaces of both systems. To the first question, "how satisfied with the interaction are you", subjects were very clearly more at ease with *Pandora* as indicated by the difference in means in table 1 [*Pandora*: median=4 mode=4 | *Last.fm*: median=4 mode=4]. If we make a small approximation and consider the data as non-constrained to the 1-5 scale, we can compute a RM-Anova on the data. This analysis shows that the difference in means is significant ($p<0.01$). Globally, users were satisfied with both systems as in both cases more than half of the subjects expressed a preference that was above the average score of the five-point Likert scale. However a solid 22.7% more users found *Pandora* excellent and in total 30% more users found it's interaction above the average mark. Strikingly, only 5% found it bellow average, against 17% for *Last.fm*.

Subjects were questioned on what had worsened their satisfaction. *Pandora* users indicated two main reasons that were that feedback options were not detailed enough (3 users), and that there was no way of having multiple artists for one radio channel[4] (4 users). For *Last.fm*, the two main problems were installation difficulties (initial effort), and interface difficulties (not intuitive, not clear or not comfortable). The difference between the two systems is striking. *Last.fm* users mentioned fundamental usability problems which have complicated the usage of the system, whereas *Pandora* users talked about some secondary issues, not fundamental in making the radio work. Furthermore, for the first system only 7 people mentioned these issues, against 16 in the second case.

Significant results	Mean (Std. Dev.)	
	Pandora.com	*Last.fm*
How satisfied with the interaction are you? ($p<0.05$)	4.1 (1.0)	3.4 (1.1)
Not significant		
Did you find it easy to provide feedback? ($p>0.05$)	3.7 (1.0)	3.5 (1.0)

Table 1: Interface quality results

A certain number of other elements were considered for providing further ex-

[4]This is not true, but clearly users did not find how to do it.

planation to this satisfaction difference, the first being *feedback*. The question "did you find it easy to provide feedback" doesn't help explain this difference. Indeed, users find both systems equally easy to operate for this topic [*Pandora*: median=4 mode=4 | *Last.fm*: median=4 mode=4]. An Anova shows us that the results are not significantly different (p=0.228). The same is true from the evaluation results of the first exposure, such that there can't be any influence of the order in which the systems are tested (p=0.464). These results are not very surprising as both radio interfaces use similar and simple systems for providing feedback, whereby the user can make a single click to show that he loves or hates a song.

More feedback issues were investigated through a proposed selection of three reasons for feeling that the user could have discovered more music. Users were asked to select a label for each of the three suggested causes. The labels were "small", "medium" and "big" problem for the three reasons: 1) feedback options being too limited, 2) hearing certain songs twice and 3) having enough time to listen. Graph on figure 3 shows the results for both systems. Clearly there is some similarity between the results. The most visible difference appears to be that more *Last.fm* users found that limitations in feedback options were a "big problem", and less so for *Pandora*. However this difference is relatively small. We believe that the feedback contrast might be an indication that *Last.fm* users did not feel that their feedback had sufficiently changed the music proposed. This is supported by the fact that both systems propose very similar actions for providing feedback. We therefore see no "a priori" reason for such a difference. Another small difference is that less *Pandora* users found the time limit to be a problem: we believe this small effect might be due to the ease of use of the interface, which just starts in a couple of seconds, even for a novice user. More results later in this paper explore this concept as well.

The interface quality questions seem to indicate that *Pandora*'s ease of use makes it a more satisfying interface than *Last.fm*. The tools proposed for giving feedback are clearly easy to use, but some users obviously would like to give more detailed indications than just "I like" or "I don't like", possibly because they felt that their feedback did not influence the proposed music sufficiently.

6.2.3 Subjective attitudes

The results on subjective questions are shown in table 2. The first subjective question was "How enjoyable were the recommended songs?". As the distributions on graph of figure 4 highlights, a strong number of subjects gave *Pandora*

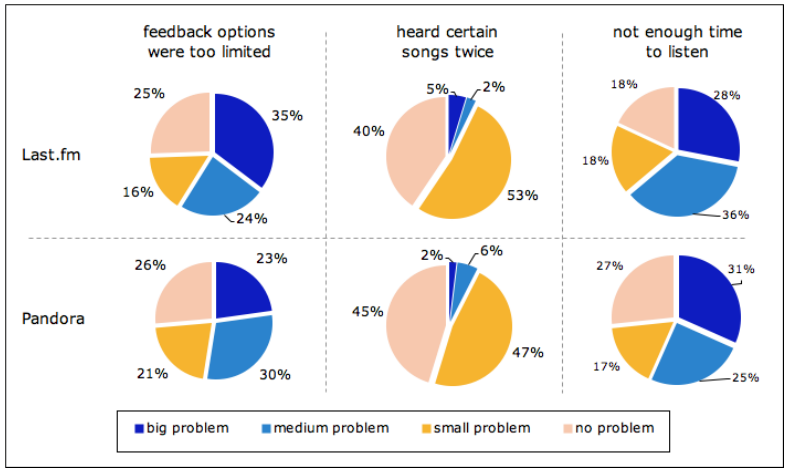

Figure 3: Graph of users' evaluation of reasons for feeling that they could have discovered more.

a score of 4 out of 5, whereas the distribution for *Last.fm* is more centered, a bit like a gaussian distribution [*Pandora*: median=4 mode=4 | *Last.fm*: median=3 mode=3]. However, a within-subject Anova shows that results are only moderately significant (p=0.08). It is interesting to observe that 67% of users gave *Pandora* an above average score (4 or 5) against only 45.3% for *Last.fm*. If we reclassify the data into three categories, *above average* (4 or 5), *average* (3) and *below average* (2 or 1), an Anova points out that the difference between systems is then significant (p=0.01). We therefore believe that there is a higher level of *global* enjoyability for *Pandora*.

The questionnaire gave the subjects the possibility to explain what hampered their listening experience in terms of enjoyability. For both systems, the main reason mentioned was the poor quality of recommended songs as eleven users of *Pandora*, respectively nineteen of *Last.fm*, reported this as the main enjoyability problem. It seems obvious that a RS cannot always be "right", and this 11-19 ratio seems reasonable. However the difference is significant ($p<0.01$) and tends to indicate that the first system provides better recommendation quality (under the constraints of the experiment setup); this result will be addressed further in

this paper. Another reason indicated by *Pandora* users, was that upon entering marginal artists or songs as a starting point, the system didn't seem to have any such "data" on which to make recommendations. This problem was reported by four users and is a known issue with such systems, especially since each song has to be analyzed and classified according to multiple attributes. Finally, a third main concern was expressed about proposed music sometimes being too similar, although all those who mentioned this point added that it was probably normal since it was the goal of the system. Including diversity in recommendations is a well know issue that many papers study [9]. *Last.fm* users also felt that marginal and obscure artists were a problem for the system (five subjects). Many other issues were mentioned for this system, each time only by one or two subjects, but nothing significant. Critiques go from "choice is too wide", to "the lyrics are missing", and mention the cold start problem "initial songs proposed were bad" or even "recommendations got worse over time", for example. *Pandora* subjects also referred to the these last two defaults, and some hinted at "too much commercial influence" or "feedback options were too limited".

Significant results	Mean (Std. Dev.)	
	Pandora.com	*Last.fm*
Was the system good compared to recommendations you may receive from a friend? ($p<0.05$)	3.4 (0.8)	2.9 (1.0)
Moderately significant		
How enjoyable were the recommended songs? ($p<0.1$)	3.6 (0.9)	3.3 (1.1)
Not significant		
The system gave more personalized recommendations based on my feedback ($p>0.05$)	2.8 (1.0)	2.7 (0.9)

Table 2: Subjective variables results

These results give us a first indication that under this precise setup, *Pandora*'s recommendations might be better than *Last.fm*'s. They also tend to indicate that *Pandora*'s interface is easier to use and corresponds better to the users' mental

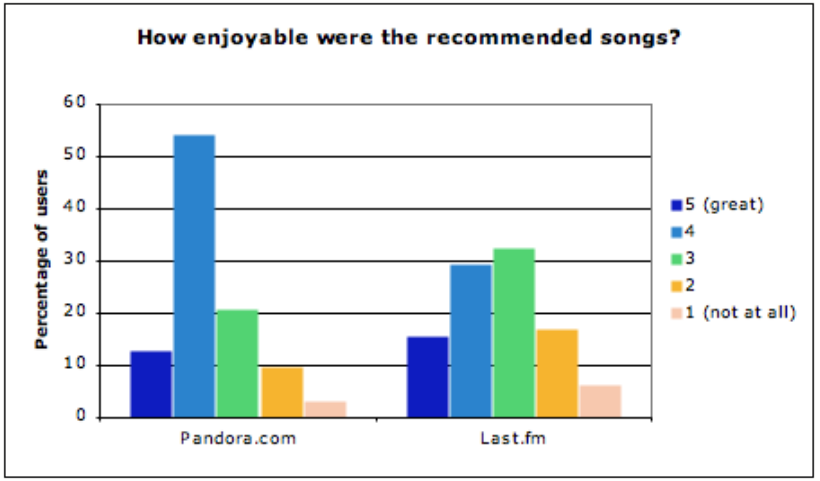

Figure 4: Graph of enjoyability of recommended songs.

models, thus making their musical experience better. However, the enjoyability measure is not so clear-cut between the two systems in terms of satisfaction: we believe that this is inherent to the music domain where any song, even randomly chosen, has a reasonable chance of being pleasing to the ears of the average listener. The following paragraph reinforces this statement.

The next subjective interrogation challenged participants to decide if they appreciated or really discovered music. The four possible answers were: "neither", "appreciate", "discover" and "both". Figure 5 shows the distribution of users' answers. It is striking to see that many more users both discover and appreciate songs suggested by *Pandora* rather than *Last.fm* and that this difference is significant ($p=0.049$). When considering the total number of subjects who selected "discover", is appears that *Pandora* is significantly better ($p=0.058$) than *Last.fm*. In other words, *Pandora* seems to not only provide more new songs, but also new songs that people like. We believe this to be an important result.

One essential point of this study was to measure the quality in terms of perceived accuracy of the recommendations. In order to do so, the students were asked "Was the system good compared to recommendations you may receive from a friend?" Testers of *Last.fm* indicated a slightly negative emphasis, as their

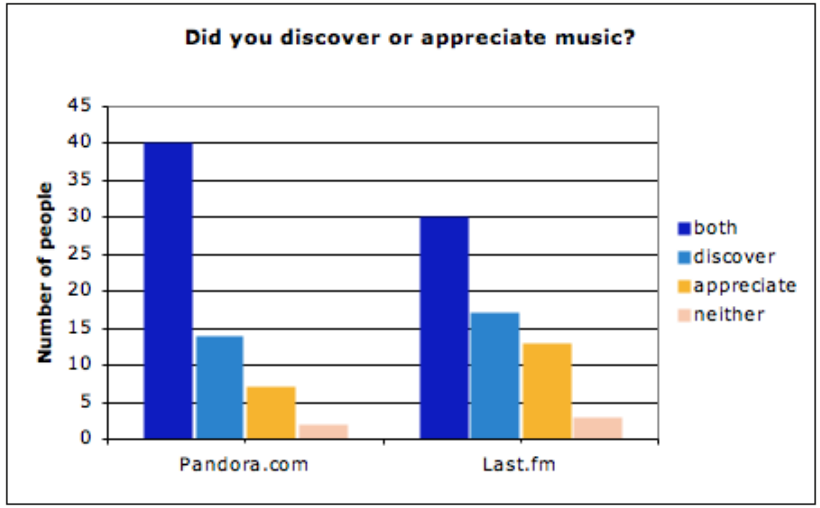

Figure 5: Graph comparing discovery and appreciation of music.

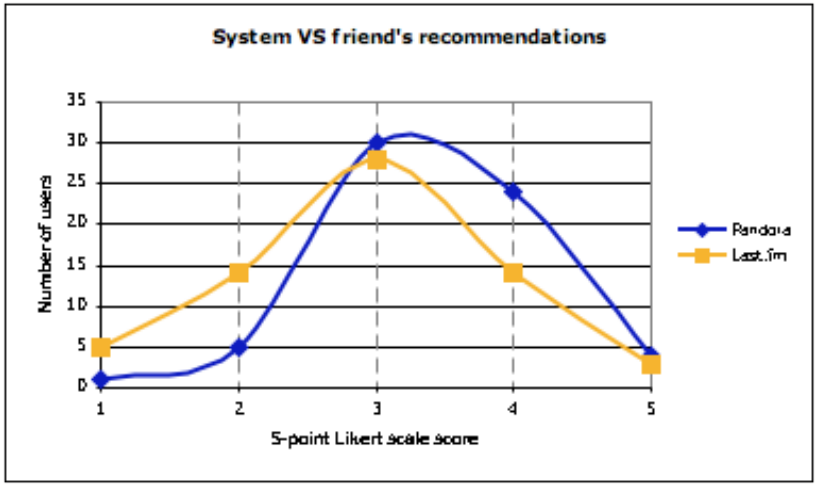

Figure 6: Distribution of appreciation of recommendations VS friends' recommendations.

average was below the middle score on the 5 point scale. On the contrary, users felt that *Pandora* was better than this middle score. [*Pandora*: mean=3.4 median=3 mode=3 stddev=0.8 | *Last.fm*: mean=2.9 median=3 mode=3 stddev=1.0]. Both medians and modes are the same which could lead to the conclusion that the difference is not significant. But a test using Anova confirmed that the difference in means is significant ($p<0.01$). The data is reported in table 2 and shown on the graph of figure 6. A trend-line has been added to facilitate the visualization of the data distributions for both systems. *Last.fm* users are clearly very centralized as in a Gaussian distribution, whereas *Pandora* users have a stronger concentration above the middle-score mark. We believe the significant separation in data frequency is an important measure because comparing system and user's (here a friend) recommendations is an indirect measure of the accuracy of the recommender system.

6.2.4 Objective quality

The objective variables were aimed at obtaining impartial measures of what users listened to, and how good systems and their recommendations were. Users were asked how many songs they listened to, or how many songs they really loved. The templates that we collected gave precise indications on how many songs students listened to, for how long they listened to the music, which songs they loved and which they hated, if a song was new to them and if they were prepared to buy it. The duration of listening was used to adjust results on a linear scale so that all results were comparable on a one hour listening period.

Table 3 shows the results from the templates. The difference between the two systems for the number of songs a user is able to listen to in one hour is rather small as the averages of 20.1 and 20.4 demonstrate. The results are not significantly different ($p>0.1$) and the standard deviation for *Pandora* and *Last.fm* are respectively 4.0 and 7.8, which are high values. This result does not surprise us: as long as the songs played do not displease the users, they will listen to roughly the same number of songs in a fixed amount of time.

Average nb of songs, in 1 hour

Nb of songs...	Median (Mean)	
	Pandora	*Last.fm*
people listened to	19 (20.1)	20 (20.4)
people LOVE	12 (12.5)	10 (9.8)
people HATE	5.3 (5.3)	6 (7.5)
people find NEW	13 (14.1)	12 (11.1)

Table 3: Objective measures from templates

The next attribute considered was the number of songs subjects really *loved*. The results show a stark difference between the two systems, in favor of *Pandora*, where on median people liked two more songs per hour than with *Last.fm*. The results are significant (p=0.021) and even more evident since the mode for *Pandora* is 10, compared to 6 for the other system.

The next attribute was about novelty. Under the current steup, *Last.fm* seems not to be performing as well since it's median value was one song fewer and the average number of songs played was three lower than *Pandora*. [*Pandora*: mode=10 stddev=5.04 | *Last.fm*: mode=6 stddev=3.65]. Again the results are significant (p=0.022).

The fourth main evaluated attribute was how many bad recommendations were made so that users came to *hate* a song. On average over one hour, *Last.fm* users hated two more songs than those proposed by *Pandora* . [*Pandora*: mode=6 stddev=3.17 | *Last.fm*: mode=4 stddev=5.60]. Results are somewhat significant according to the p-value obtained (p=0.061). A deeper analysis of the results shows that *Last.fm* has more records of users hating a high number of songs and that these cases do not come from adjusted values, but are all from people who had listened for a full hour.

The templates were also used to evaluate how many songs the subjects were ultimately prepared to buy. As could be expected, the most frequent answer is "0". However, a small number of people did show some interest in purchasing some of the music. If we only consider the first exposure to the two recommender systems, we can see that out of the 64 candidates, 29 (i.e. 45%) said they would be ready to purchase a song online, with the median value of the non-zero answers being 2. The difference between the two radio systems when measured in this context is not significant.

6.2.5 User preferences in systems

The final part of the study concentrated on obtaining the users' final preference. This was done through a set of six related questions where the students had to choose between the two systems for each question. The goal was to differentiate several dimensions which can each play an important role in a recommender system. The results presented are significantly in favor of *Pandora* (see table 4).

Questions	Nb people	
	Pandora	*Last.fm*
Which system do you prefer most?	71%	29%
Which interface do you prefer for getting music recommendation?	70%	30%
Which interface de you prefer to use as an internet radio?	62%	38%
Which interface inspires more confidence in you in terms of its recommendation technology?	70%	30%
If I want a recommendation in the future, I will be likely to use:	66%	33%
I felt comfortable using the following interface:	66%	33%

Table 4: Users' preference in systems

The first preference question asked subjects directly which system they preferred most, and the answer was very clear as 71% users preferred *Pandora*. A Chi-Square test of independence was computed (for all preference questions) to make sure that these results were not influenced by the order in which the students tested the systems, and the test is conclusive that there is no order-effect correlation since the p-value is high (p=0.57).

Further preference results were just as conclusive. When asked what interface users preferred in terms of music recommendations, 70% voted for *Pandora*. Again tests of independence show no order influence (p=0.35). And to the question "Which interface do you prefer to use as an internet radio?", subjects had strong opinions again as only 38% voted for *Last.fm* (p=0.16). However, this result is not as clear-cut as the previous ones. Curiously, this happens on the question where a more social dimension is approached, an internet radio. Sub-

jects previously indicated, very clearly, that *Pandora* had a better interface and was easier to use, two aspects clearly important for an internet radio. Despite that, they still seem to indicate that this is the best function for *Last.fm*: being an internet radio, in the more classical way. We believe this possibly comes from the vast amount of social manipulations that can be done on *Last.fm*'s website, such as writing blog entries, defining musical friends, leaving comments and many more similar actions.

One of the most important dimensions in these preference questions was to determine the quality of the recommendations. So we asked the students: "Which interface inspires you more confidence in terms of its recommendation technology?". 70% clearly designed *Pandora* as the most accurate system and ordering has no effect (p=0.47).

To the question "Which system would you use to get a recommendation in the future", participants designated *Pandora* in 66% of cases; ordering has no effect (p=0.45). The last preference question considers the interface design, one last time, through the word "comfort". Again, students vote for *Pandora* in 66% of cases; ordering has no effect (p=0.32).

We extended the analysis of the preference questions' results by computing the inter rater-reliability amongst the answers. This is useful for determining how much homogeneity there is in the answers given by the users, therefore indicating if these six preference questions were perceived as representing different dimensions or not. The computed Intra-class correlation coefficient, ICC = 0.29, shows that there is no consensus across the questions.

Users' answers for the preference questions are highly in favor of *Pandora*. Whether considering the recommendation interface or simply the best system, the main trend of responses always points to *Pandora*. Furthermore this 30%-70% separation does not come from two groups of people voting exclusively for one system, but reflects user's diverse opinions on multiple criteria used to judge these two music RS.

6.3 Correlation Analysis

Correlation analysis (table 5) among the measured variables shows that enjoyability of songs, interface satisfaction, and the number of songs loved are the most important factors in predicting the relative quality of recommendations as being better than what the user may get from their friends. Interestingly but not so surprisingly, the number of songs subjects were prepared to buy correlates positively with recommendation quality. Analysis of users' detailed comments show that the main problems causing users dissatisfaction with *Last.fm* 's interface are the initial time required to set up the proper environment (initial effort), the time it takes to get useful recommendation (time to recommendation) and the fact that the interface is not intuitive and comfortable to use (simplicity). However it is not clear whether the high time to recommendation is a problem as such, or if it is above all linked to the other two dimensions as a kind of side effect. As for the shortcomings of *Pandora*, users wished that they could provide more refined feedback. Another deficient feature for *Pandora* was that users didn't always find how to create a radio channel with more than one artist.

Factors that predict RS quality	*Corr. (sig.)*
Enjoyability of recommendations	0.760 (0.000)
Interface satisfaction	0.574 (0.000)
No of songs subjects loved	0.315 (0.000)
No of songs subjects were prepared to buy	0.289 (0.001)

Factors that do not predict RS quality	*Corr. (sig.)*
No of songs people listened to	-0.062 (0.484)
Interrupted whilst listening	0.019 (0.830)
Trying other features whilst listening	0.035 (0.697)
Would you have discovered this system on your own?	0.127 (0.152)

Table 5: Prediction quality of recommendations

7 Conclusion and Future Work

Two music recommender systems were compared side-by-side in a within-subject user study involving 64 participants. Each participant evaluated one system first, and then the other a week later. Thus, a total of 128 evaluations were performed. The main goal was to investigate the adoption of recommender technology and users subjective perception of the respective systems as new users. For this objective, we chose to compare *Pandora*, a content-based system, with *Last.fm*, a social or collaborative filtering system.

The study results show that users significantly prefer *Pandora* to *Last.fm* as a general recommender system, are more likely to use *Pandora* again, prefer to use *Pandora*'s interface for getting music recommendations and as an internet radio, and perceive *Pandora*'s interface as more capable of inspiring confidence in terms of its recommendation technology. Moreover, users are generally more satisfied with *Pandora*'s interface, and found the songs that it suggested were significantly more enjoyable and perceivably better than their friends' suggestions. Finally, under this specific setup, users also loved more songs suggested by *Pandora* than *Last.fm*, found the songs more novel, and disliked fewer of the suggested songs from *Pandora*, albeit this might be directly linked to *Last.fm*'s inability to recalculate users' profiles in less than a few days.

Our evaluation of *Pandora* and *Last.fm* provides a first understanding of how recommender websites attract new users as a result of the site design. Based on our result analysis, we are able to derive a set of general design principles which can also be applied to other domains such as movies and travel products: 1) minimizing user effort such as the time to register, download and get recommendations, 2) maximizing the quality of recommendation such as accuracy relative to a commonly shared measure (e.g., to friends' suggestions), such as to maximise enjoyability and novelty, and 3) maximising the interface's ease of use. According to our study, users clearly prefer such recommender systems and as a result are more convinced of its underlying technology. This finding is consistent with the fact that *Pandora* was voted by Time magazine as among the top 50 "coolest websites", among only 8 other entertainment websites.[5] As simple as they may appear, these initial findings show that focusing on the recommendation's technology alone is not enough to attract new users. An analysis of the website design and especially the human factor aspects are crucial in understanding users' technology adoption issues. Furthermore, the captured dimensions are highly similar to those highlighted by repeated studies by Forrester Research, such as [11], which stress that most important factors in user-web-interaction are ease of use (e.g., minimising user effort) and content quality (e..g., maximising recommendation quality).

Even though the direct intention to purchase music at both sites was not very significant, users expressed more intention to return to *Pandora* . An increased intention to return due to positive experiences gained on the initial visits is likely to bring revenue for the websites. According to a recent marketing report by WebSideStory, returning visitors are 8 times more likely to purchase than first-time visitors.

In a long-term perspective, we aim to study not only adoption issues, but also a wider range of user experience aspects including user loyalty. In the second planned user study, we will focus on the elements composing the dimensions highlighted by this study, and explore the social features that recommender systems offer and evaluate to what extent they provide motivations to attract users and keep them there as loyal costumers. We hence hope to enrich and refine the design guidelines. We also plan to investigate how users' moods influence the perceived accuracy and enjoyability of recommendations.

[5] See Time Magazines July 26^{th} 2006 online issue.

References

[1] ADOMAVICIUS, G., AND TUZHILIN, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on 17*, 6 (2005), 734–749.

[2] BALABANOVIC, M., AND SHOHAM, Y. Combining content-based and collaborative recommendation. *Communications of the ACM 40*, 3 (March 1997).

[3] BONHARD, P., HARRIES, C., MCCARTHY, J., AND SASSE, A. M. Accounting for taste: using profile similarity to improve recommender systems. In *Proc. CHI '06* (2006), ACM Press, pp. 1057–1066.

[4] BREESE, J. S., HECKERMAN, D., AND KADIE, C. Empirical analysis of predictive algorithms for collaborative filtering. pp. 43–52.

[5] BURKE, R. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction 12*, 4 (November 2002), 331–370.

[6] CHEN, L., AND PU, P. Evaluating critiquing-based recommender agents. In *AAAI* (2006).

[7] GOOD, N., SCHAFER, B. J., KONSTAN, J. A., BORCHERS, A., SARWAR, B., HERLOCKER, J., AND RIEDL, J. Combining collaborative filtering with personal agents for better recommendations. In *AAAI '99/IAAI '99* (1999), pp. 439–446.

[8] HERLOCKER, J. L., KONSTAN, J. A., AND RIEDL, J. Explaining collaborative filtering recommendations. *CSCW'00* (2000), 241–250.

[9] MCGINTY, L., AND SMYTH, B. On the role of diversity in conversational recommender systems. In *ICCBR* (2003), pp. 276–290.

[10] MCNEE, S. M., LAM, S. K., KONSTAN, J. A., AND RIEDL, J. Interfaces for eliciting new user preferences in recommender systems. In *User Modeling* (2003), pp. 178–187.

[11] MOIRA DORSAY, HARLEY MANNING, C. L. C. Death by a thousand cuts kills web experience. Tech. rep., Forrester Research, Inc., August 2006.

[12] PENNOCK, D., HORVITZ, E., LAWRENCE, S., AND GILES, C. L. Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach. In *Proc. UAI 2000*, pp. 473–480.

[13] PU, P., AND FALTINGS, B. Enriching buyers' experiences: the smartclient approach. In *Proc. CHI '00* (2000), ACM Press, pp. 289–296.

[14] PU, P., VIAPPIANI, P., AND FALTINGS, B. Increasing user decision accuracy using suggestions. In *Proc. CHI '06* (2006), ACM Press, pp. 121–130.

[15] RASHID, A., ALBERT, I., COSLEY, D., LAM, S., MCNEE, S., KONSTAN, J., AND RIEDL, J. Getting to know you: Learning new user preferences in recommender systems. In *Proc. IUI 2002*, pp. 127–134.

[16] RESNICK, P., IACOVOU, N., SUCHAK, M., BERGSTORM, P., AND RIEDL, J. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proc. of ACM CSCW'04*, ACM, pp. 175–186.

[17] SCHAFER, J. B., KONSTAN, J. A., AND RIEDL, J. Recommender systems in e-commerce. In *ACM Conference on Electronic Commerce* (1999), pp. 158–166.

[18] SINHA, R. R., AND SWEARINGEN, K. Comparing recommendations made by online systems and friends. In *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries* (2001).

[19] SWEARINGEN, K., AND SINHA, R. Beyond algorithms: An hci perspective on recommender systems, 2001.

[20] SWEARINGEN, K., AND SINHA, R. Interaction design for recommender systems. *Designing Interactive Systems* (2002).

[21] VIAPPIANI, P., FALTINGS, B., AND PU, P. Preference-based search using example-critiquing with suggestions, To appear in the Journal of Artificial Intelligence Research, 2006.

[22] ZIEGLER, C.-N., MCNEE, S. M., KONSTAN, J. A., AND LAUSEN, G. Improving recommendation lists through topic diversification. In *Proc. WWW '05*, ACM Press, pp. 22–32.

Nicolas Jones obtained his master's degree in Computer Science in 2005 at EPFL, the Swiss Federal Institute of Technology in Lausanne. In parallel with his studies he co-created an Internet Services company, were he pursued his interest for modern and novel interaction mechanisms. Following his diploma project at IBM Zrich Research Laboratory, he is currently carrying out his PhD in the Human Computer Interaction's Group at EPFL under the supervision of Dr. P. Pu. Working on multiple topics regarding Recommender Systems, the primary focus is on user issues in entertainment recommender systems (i.e. music, movies.) which lead to the acceptance and adoption of the system, and the impact of implicit and explicit preference elicitation techniques.