User Acceptance Issues in Music Recommender Systems

EPFL Technical Report HCI-REPORT-2009-001.

Nicolas Jones
Human Computer Interaction Group
Swiss Federal Institute of Technology
nicolas.jones@epfl.ch

Pearl Pu
Human Computer Interaction Group
Swiss Federal Institute of Technology
pearl.pu@epfl.ch

ABSTRACT

Two music recommender systems were compared side-by-side in an in-depth between-subject lab study. The main objectives were to investigate users' acceptance of music recommendations and to probe the main technology acceptance model in the environment of low involvement recommendations. Our results show that perceived usefulness (quality) and perceived ease of use (effort) are the key dimensions which are sufficient to incite users to accept recommendations, and that the adapted model is suitable for entertainment recommenders. Measures of quality such as accuracy, enjoyment, satisfaction and having music tailored to a user's taste are directly correlated with acceptance, and measures of effort like the initial time to reach interesting recommendations and the ease of use for discovering music are strongly linked to acceptance. The study shows how important it is for a music recommender system to take into account users' emotions and mood. Finally, the results highlight the necessity for low-involvement recommenders to be highly reactive.

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces—*Evaluation/methodology*.

General Terms

Human Factors, Performance, Experimentation.

Author Keywords

Recommender systems, interaction design, usability evaluation, acceptance issues, recommendation technology, entertainment website evaluation.

INTRODUCTION

The year 2000 saw the climax of the first dot-com bubble. This high-speed internet revolution catalysed wide-spread, global and fundamental changes in the way people use their computer, communicate, make business and strongly influenced the present society. Today in 2007 a similar ambience is brewing on the web. Thousands of new websites appear every hour and new services are launched every day. The competition amongst websites of similar interest is ongoing and those who don't stay up-to date risk loosing their audience in a very short time. In this day and age of the ever-changing web-fresco, recent trends have seen the emergence of personalised services, where users must create a profile on each website and maintain a regular input to benefit from these customised services. Rapidly users tend to become overwhelmed by the profusion of possibilities and end-up using only a small subset of what is available. It is in this context that this paper explores the mechanisms which lead users to accept suggestions from a system, and how they resort to adopting one system rather than the other.

Technology acceptance research investigates how and when users come to accept and use a website, a software system, or a technology. The history of this subject can be traced to as early as the 1970s relating to the adoption of new technological innovations. With the arrival of computers, software, and lately websites, there has been a growing interest to apply the general framework of technology acceptance to specific domains. Our current research examines this issue in the particular domain of recommender systems which offer entertainment products (music, books, films, etc.). We decided to test the original technology acceptance model (TAM) proposed by Davis [2] in 1989. This model will be used in devising our research model to measure and understand users' experiences with recommender systems. The TAM suggests that when users are presented with a technology, a number of factors influence their decision about how and when they will use it, notably the perceived usefulness (PU) and the perceived ease-of-use (PEOU). PU is defined as the degree to which a person believes that using a particular system would enhance his or her job performance and PEOU as the degree to which a person believes that using a particular system would be free from effort.

User Acceptance Research for RS

Recommender systems (RS) are a recent tool used by websites to help users access the ever-growing set of products and data available on the Internet. Yet, on the scale of internet technology history, RS are not "new". Historically, recommendation technology was used in recommendation-giving sites where the system would observe a user's behavior and learn about his/her interests and tastes in the background. The user would select articles to read [11], or items to purchase, and the RS would then propose items that may potentially interest him/her 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 were offered as a value-added service for users to discover new items and as a way for the site to interest users in buying items that they did not look for initially. If the context of use of recommender systems had remained unchanged, understanding how users may *accept* recommender systems and recommendation results would not have been such a crucial field of study.

However, a trend has recently emerged where users go to websites to actively *seek* advice and suggestions for electronic products, vacation destinations, music, books, etc. They interact with such systems as first-time customers and may not get the usual benefit of receiving recommendations "automatically" [12] due to the lack of a personal preference history. Compared to receiving unsolicited recommendations, users who specifically seek recommendations are likely to have a higher level of expectations on the results they obtain and the ease of use of the system. In the seeking context, a user makes a conscious decision to use a system for a specific goal. If they do not find what they are looking for, or the system is too hard to use, they may quickly leave. We therefore argue that as recommender systems are broadening their scope of use, the acceptance process is becoming more complex and pertinent to study. Site designers must somehow identify the right balance between the benefits they offer relative to the effort they require from users in order to increase their ability to attract new users and provide a high level of staying power.

Motivation

We specifically chose music recommender systems to conduct our research based on the ease of accessibility to such systems. Music is an attractive and highly inspiring topic where it is easy to motivate users to be involved in a trial compared to other domains such as news articles. Furthermore, music has a relatively short validation process to determine the quality of recommendation results. Compared to books and movies, a user can more quickly and easily determine if a recommended song is enjoyable, novel, etc. We also decided to initially focus on low-involvement entertainment products because they carry a smaller financial commitment compared to most of the other entertainment-related commodities, such as electronic and travel products. Music items are available in large quantities and in a wide range of varieties. They do however propose other challenges: because they are relatively easy to acquire, users are unlikely to spend much time choosing them, hence the name low-involvement products.

The TAM is an important model, which has become very influential over time. Today's challenges often evolve around building a website which users will accept and ultimately adopt. Yet this still remains a complex task, where only a few succeed. With the growing popularity that RS are getting, it is important to work on better understanding the required fundamentals when building a RS. In this perspective, testing the TAM on recommender systems is important.

We selected the TAM as our baseline for two main reasons. First of all, as the section on related work later explains, numerous improvements and variations of the original model have been proposed and tested, but each-time the same core elements remain essential. Secondly the TAM is extremely simple, yet fits a high number of situations. The combination of these two reasons lead us to select it. Our research work was planned in several stages. The first main challenge was to identify and validate the parameters used to measure the three aspects of the TAM in the specific domain of recommender systems: the perceived usefulness, the perceived ease of use, and the acceptance of such systems. Although some previous research investigated user experience issues involved in recommender systems, most of the variables were derived from our own interviews with subjects in pilot studies. There are three main groups of variables for perceived usefulness: 1) the entertainment value provided by the recommended results such as the enjoyability, novelty, satisfaction, and accuracy of the suggested songs 2) the system's ability to adapt to user feedback and 3) the completeness of the system's database. For the ease-of-use aspect, we decided to measure both perceived ease of use and the actual user effort in terms of the time to sign up and the time to recommendation (the time between signing up and the time that a user starts to enjoy the recommended songs). For the acceptance aspect, we divided our parameters into the acceptance of recommendation results, the acceptance and adoption of the system, and users' intention to use the system again.

This classification led us to design and conduct an initial in-depth with-in subject user study in March 2006 involving two music recommender systems, *Last.fm* and *Pandora*. We were very interested in understanding how users accept a specific recommender system as a result of the recommender technology used and the surrounding features. The conclusion of that study suggests 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 (see [5]).

We decided to expand upon this initial study by providing the users with an opportunity to test the systems over a longer period of time. Originally, only a single one-hour session was used for users to interact with each system, constraining the measure of recommendation results, since more time was needed to produce personalized recommendations. We also took this opportunity to conduct detailed interviews. We wanted to gather information and explanations on why a user more easily accepted one system than the other, and further explore several dimensions left "open" by the initial study. Finally, recent developments of *Last.fm* mean that users can finally listen to music just like *Pandora* users, in their browser and without having to install an application. This change made both systems easier to compare. For these reasons, we decided to conduct a new study, this time as a between-group lab study, with two sessions separated by a fifteen-day time interval. 20 novice users were selected for testing the two systems, their entire interaction process were recorded and each session was concluded by a short interview. To conclude the whole experiment, participants had to complete a survey of 28 questions. Due to the lengthy nature of the study, we decided not to perform a with-in subject study because of the potential fatigue risk that such experi-

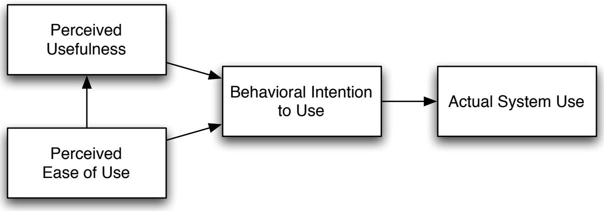


Figure 1. The Technology Acceptance Model (TAM).

ment would impose on our subjects.

The rest of the paper presents the findings of our second user study, and is organized as follows. We first present related work on the issue of technology acceptance. The paper then shortly introduces the two music recommender systems being tested, before presenting our research model, based on the TAM. The experiment description, with the exposition of the evaluation framework, the procedure used for the experiment and the set of questions used in this study follow. Then we explain the results of our user study, before concluding and discussing future work.

BACKGROUND AND RELATED WORK

Since the focus of this paper is on user-related issues and since a comprehensive development on technical aspects was made in our initial paper, the two main technologies of recommendation will not be presented here. On the contrary, this section will highlight past work on elements leading to *acceptance* and will start by presenting Davis et al.'s Technology Acceptance Model.

Technology Acceptance

Computer technology acceptance by users is a topic which Davis et al. tackled already back in 1986, forging research work which has become very influential in the field today [2, 3, 13]. He introduced the Technology Acceptance Model (TAM), figure 1, an adaptation of the Ajzen and Fishbein's Theory of Reasoned Action, which hypothesizes that perceived usefulness and perceived ease of use influence a user's intention to use a system and eventually how he will use it. The factors factors are defined as:

Perceived usefulness (PU) is defined as the degree to which a person believes that using a particular system would enhance his or her job performance.

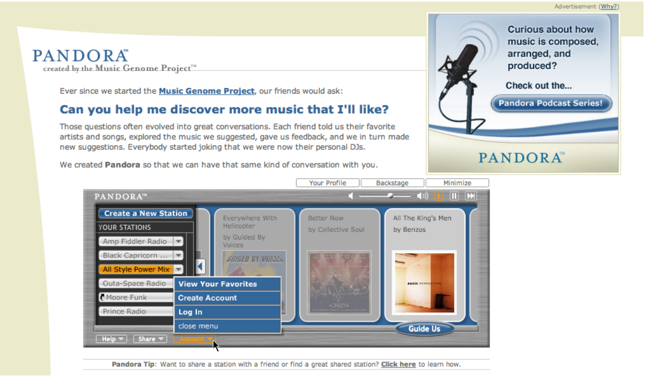
Perceived ease-of-use (PEOU) is defined as the degree to which a person believes that using a particular system would be free from effort.

These two dimensions, PEOU and PU, have since then been at the heart of the research on user related issues that lead to acceptance and eventually adoption. Recently, Hassanein and Head wrote about building trust through socially rich web interfaces [4] (i.e. e-commerce websites) in a study where the TAM was used. Indeed, trust in an online shopping context is a complex issue which deals with a broad range of aspects. In a push to explore trust and its determinants, the research coupled the TAM with "social presence" and "enjoyment" as an interconnected network leading to trust. Previously, Koufaris and Hampton-Sosa had set the groundwork for similar research by examining the role of the experience with the website in customer trust online [8]. The TAM was also used as a baseline but augmented with the influence of "enjoyment" and "perceived control" on PU and PEOU, and linking the whole model to effects on "intention to return" and "intention to purchase". In both studies, results reinforced the importance of the TAM, while highlighting the fact that other dimensions belong to the model, often as a catalyzers of either PU or PEOU or both.

A good example of work on these potential components is [6], where Kamis and Stohr studied parametric search engines. The goal was to model the effectiveness of four parametric search engines in using search effort and domain knowledge to increase decision quality, decision confidence, PEOU and PU. Based on their previous work, they came up with a two-dimensional classification where subjective and objective elements were placed with respect to decision inputs and decision outcomes. This way they formed a research framework which encompassed factors like search effort, domain knowledge, decision quality, decision confidence, PEOU and PU. Based on this, they created a model where search effort and domain knowledge influences decision quality, which in return impacts decision confidence, itself being directly linked to PU and PEOU. This research showed that search efforts and domain knowledge were mediated through decision quality and decision confidence, and that these impacted both PEOU and PU.

However, not all models are as succinct as the TAM. Recently at the CHI 2006 conference, McNee, Riedl and Konstan proposed an analytic model for RS, which they named Human-Recommender Interaction (HRI) [10]. They proposed this framework and methodology because they felt the need to obtain a deeper understanding of users and their information seeking tasks. Their model is not based on the TAM, it rests on three pillars of what they call the interaction process: the recommendation dialogue, the recommender's personality (perceived by the user over time) and the user's information seeking task. For each pillar, they come up with a range of dimensions that are thought to influence users' behavior. In the case of our music RS study, we will mainly be observing the user's interaction, part which in the HRI model is classified as the Recommendation Dialogue. The later is divided into eight elements such as correctness, quantity, saliency, etc. Two elements, usefulness or usability, can be considered as equivalent to our definition of PU and PEOU. This HRI model suggests that these two dimensions are far from being enough to model the main user-interaction dialogue and that they stand parallel to six other dimensions such as serendipity.

The main results of all these studies on users' acceptance of technology and recommendations show the importance of very domain specific dimensions. However, and to the exception of the HRI model, the perceived ease of use and per-



Ever since we started the Music Genome Project, our friends would ask:

Can you help me discover more music that I'll like?

Those questions often evolved into great conversations. Each friend told us their favorite artists and songs, explored the music we suggested, gave us feedback, and we in turn made new suggestions. Everybody started joking that we were now their personal DJs.

We created **Pandora** so that we can have that same kind of conversation with you.

Pandora Tip: Want to share a station with a friend or find a great shared station? Click here to learn how.

Figure 2. A snapshot of Pandora's main GUI with the embedded flash music player.

ceived usefulness stick out as key features. We hence decide to experiment with the TAM in this low-involvement environment of music recommendations. PEOU and PU are our two main pillars, then decomposed into smaller dimensions. Our model is explained in Research Model section.

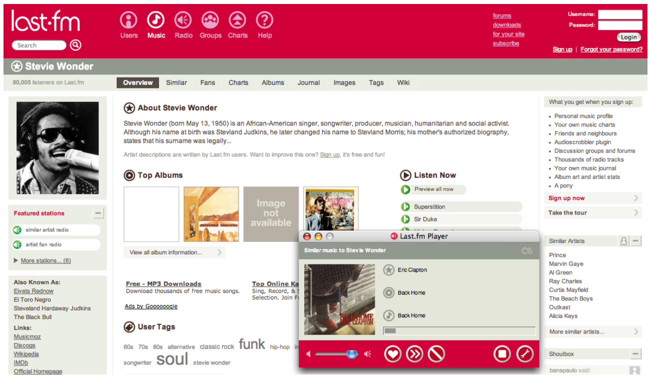
THE TWO MUSIC SYSTEMS

Pandora.com

When a new user first visits *Pandora* (figure 2), a flash-based radio station is installed within 10-20 seconds. Without any registration requirement, 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 recommends to you next. You can create 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, while 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*, it appears that this is an example critiquing-based recommender, based on users' explicitly stated preferences. According to information published on its website, *Pandora* employs professional musicians to encode each song in their database into a vector of hundreds of 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. At the time of the study, the system included songs from more than 10'000 artists and had created more than 13 million stations.

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



Stevie Wonder (born May 13, 1950) is an African-American singer, songwriter, producer, musician, humanitarian and social activist. Although his name at birth was Stevland Judkins, he later changed his name to Stevland Morris; his mother's authorized biography, states that his surname was legally...

Artist descriptions are written by Last.fm users. Want to improve this one? Sign up, it's free and fun!

Free - MP3 Downloads
Download thousands of free music songs.

Top Online Ka
Sing, Record, & S
Selection. Join Fr

Ads by Goooooogle

Figure 3. A snapshot of Last.fm's main GUI, with the music player application in foreground.

by matching users with people who like similar music and to generate personalized radio stations (called recommendation radios) for each person. Based on information from the website and the ways *Last.fm* behaves, we assume that it uses user-to-user collaborative filtering technology. However, it is possible that it also relies on some content-based technology in parts. It is a social recommender that knows little about the songs' inherent qualities, and functions purely based on users' rating of items and tagging.

Last.fm is a rich website that incorporates a flash music player and provides an optional plugin for recording your music profile through a conventional music player like iTunes. A user can start listening to music without necessarily having an account. However the rating functions are only enabled when the user creates an account. You can then specify an artist's name, such as "Miles Davis", and 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, "Miles Davis' similar artist radio", of songs that belong to that group, and for each song, press an "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" and later use those to launch a new radio station. Additional features are proposed on the website, as shown in figure 3. 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.

RESEARCH MODEL

The following section presents the research model that drives our study and which defines which dimensions are tested in the final questionnaire. As highlighted in the background & related work section, many dimensions which influence a user's perception of a website and his potential acceptance, are associated with the perceived usefulness and perceived ease of use initially proposed by the TAM. However this was studied in conditions different from those of low-involvement RS, where the problematic of a user's interaction is very task or context specific where several different dimensions are essential. We feel there is a need to explore *acceptance* from the basics and this is what we propose: instead of including PEOU and PU within other cat-

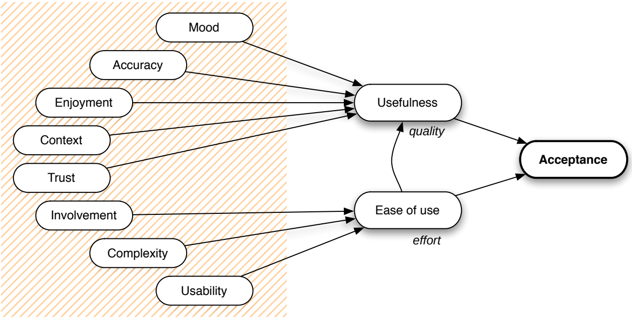


Figure 4. Dimensions of the research model.

egories such as presented in other models (for example the HRI model of [10]), we consider them as key categorizers and include several other dimensions within them, just like the TAM initially did. In music, the usefulness for a RS can be defined as the *quality* of the songs proposed. And for the ease of use of a recommender website, we are essentially speaking of *effort*. This is the core classification behind our model shown in figure 4.

So what are the features that impact PU in this field of low-risk products such as entertainment (music, books, films)? Obviously elements such as enjoyment and satisfaction are part of what defines a RS's quality, just like the context in which we listen is important. We also considered other dimensions like how diverse the songs were, if they were new to the user and if he approved of the ones he heard that he already knew, helping build his trust in the system. If the songs are tailored to the user's taste and suit his mood, this could also clearly impact his perception of the RS's usefulness.

The ease of use of a RS website can also be separated into several dimensions. The amount of involvement necessary to obtain the desired songs is one part of our model. We also consider that the complexity of the system will impact the user's perception of effort and ease of use. Complexity in itself is not a sufficient measure, as a simple website can still suffer from usability issues which can make it very challenging to operate.

We are of the opinion that these features are all foundational characteristics that support PU and PEOU, and that they lead to the user's acceptance of a RS's recommendations. The exact questions that we asked to evaluate these multiple dimensions are detailed at the end of next section.

EXPERIMENT

The experiment was conducted as a between-group in-depth lab study with 20 participants (and multiple pilot studies). An in-depth study was favored in order to first observe the users' interaction, and secondly to be able to discuss more fundamental issues that affected users. Due to the lengthy nature of the study, we decided not to perform a with-in subject study because of the potential fatigue risk that such an experiment would impose on our subjects. In order to have a balanced study which covered different types of interactions, two kinds of users were selected: the first half were computer and communication science Ph.D. students at university, and the second half were non-computer science people, at the university level and who used a computer regularly, making the range of users from normal to expert. Because of the slightly advanced nature of the topic, no beginner users were selected. A financial incentive was proposed to ensure that the participants were serious about the experiment, and all subjects took part in a draw to win a high value present. With the exception of one user, all had never heard of or used *Last.fm* and *Pandora* ensuring that the selected participants represented sound first-time user experiences. People were randomly attributed which system they would be testing in order not to introduce any bias.

Participants' Profiles

Seven of the participants were female. Most users were in the 25-30 age group, with two in the 18-24 range, and three in the 30-40 range. As a base measure of the users' affinity for music in general, we chose to ask a subjective question to assess the users' own perception of their bond with music: we questioned them about the size of their personal music collection. Most of the subjects had an average collection, except for three who thought theirs was small and three who qualified theirs as being large. We operate under the assumption that there is no bias in terms of the users' connection to music, so the uniform and centered distribution of this measure is a positive factor for the quality of our results.

Evaluation Framework and Procedure

The experiment was performed on a single machine, guaranteeing the same setup, conditions and material for each tester. A set of high quality earphones were provided allowing each subject to feel completely immersed in the listening experience and to listen at the volume of their choice. The experiment was conducted in two distinct half an hour periods, separated by a 15 day lapse. The interval was necessary since *Last.fm* requires a period of several days before it starts making personalized recommendations for new users. Although *Pandora* does not require such an interval, we imposed it to all the users to make the experience with both systems comparable. At the end of each half hour, users had a quick oral interview to verify elements such as their first impression and any potential observations. To conclude the experiment, users had to answer an online questionnaire of 28 questions (described in part) at the end of the second half hour.

Each user interaction was remotely observed through a Virtual Network Computing (VNC) client, and directly encoded into text by a unique observer. The experiment machine was connected to a high-speed internet access and the VNC was adequately configured, to ensure that no bandwidth shortages occurred. The users were informed that their interaction was being observed.

Instructions

The users taking part in this study received precise written instructions on the tasks they had to complete. The experiment was divided into three steps which are described hereafter.

Step 1 An outline of the user study was provided before asking some background questions. The outline started by informing users about the topic of the experiment (i.e. listening to music). The occasion was taken to encourage them to relax and not take this as a "test", to reduce the number of eventual outliers. Users were also informed about the experiment's unfolding in two short phases (step 2 & 3): the first to get accustomed to the system, and the second to thoroughly test it (and hopefully profit from it), the two phases being separated by the fifteen day interval. Users were informed that there would be a short online questionnaire and a small post-study interview to conclude the experiment. To conclude this step, the outline was followed by six background questions about their profile.

Step 2 The second step started by informing users of which system they would be testing. In order to make the comparison between both music recommender systems possible, users were not given detailed tasks to complete, but a scenario to follow. The goal of this part was to create an account and to get used to the system. When users felt comfortable with it, they could stop. A rapid interview wrapped up this step.

Step 3 To conclude the experiment, users were proposed another scenario where this time the goal was to get some recommendations for discovering new music in a half hour session. Once finished, users were guided to the main questionnaire, before having a final short interview.

Experiment unfolding

This subsection explains how the experiment took place and issues related to experiment bias. The experiment was carried out over one month. In this lapse of time, both services remained constant throughout the experiment, with respect to features tested by users. An administrative problem was encountered when *Pandora* changed its licensing terms, restricting access to U.S. citizens only (through the detection of the connecting IP address). At the time, five *Pandora* users hadn't yet finished the experiment. A temporary solution was found through the usage of a VPN access in America, before later getting a reply from *Pandora* and obtaining that the IP address of the experiment machine got unblocked for a week, allowing to conclude the experiment without changing the 15 day interval for each user. The only direct user problem occurred when one *Last.fm* user changed the language of the website and was subsequently not logged in anymore, preventing him from rating and tagging songs. He ultimately had to repeat the entire login process. Such a bug in the system is, to say the least, surprising for a website of over 12 million users.

After completing the pilot studies, and seeing the ease with which users followed the provided guidelines, we decided to allow users to continue to use the music recommender system on their personal computers during the fifteen day interval. This allowed us to take into account the emotional mechanisms that occur when users discover a new service and become more familiar with it.

For the *Last.fm* users, their musical profile was examined after the first session to see how many songs had been recorded to their profile by the system. This was necessary because the system only considered that a song had been *listened to* if it was heard for the shorter of three quarters of the song length or 3 minutes. The natural behavior of first time users was to skip through songs in order to better understand and explore the proposed features. This situation, detected in the pilot studies, led us to define a baseline before allowing users to come back for the second session: at least five songs must have been recorded to their profile. When this was not the case after the first session, we asked the involved user to listen to a few more songs at home, and ensure that at least five songs were saved to his/her profile.

While users were testing a music recommender system in our lab, we occasionally interrupted them to make sure they remained on-topic, in order to ensure the coherence of the experiment. We understand that the comparison of two full websites only makes sense if they propose the same features and can be used to pursue the same goal. For this reason we decided of the following interruptions, for session 1 and 2 (S1 & S2):

S1 Both systems permit the user to start listening to music without having to create an account. When a user had not created an account after 15 minutes, we interrupted.

S1 *Last.fm* provides several ways of listening to music. When a user was seen only listening to 30 second preview samples, we interrupted after 15 minutes to explain that because of radio licensing terms, full songs were only available when listening to a "radio".

S2 If after 15 minutes user's had not started their "personalized recommendation radio", we checked with them to determine whether they knew it existed, without however forcing them to use it.

Questions

The final questionnaire was composed of 26 assessment questions (refer to table 1), on a Likert scale from 1 (strongly disagree) to 5 (strongly agree). In order to keep questions balanced and natural, four questions used a reversed scale (Q21, Q23, Q26, Q27). Two textual questions, Q20 & Q22, completed the survey.

RESULTS AND ANALYSIS

In this section we report the results from our study. We first focus on the results from the questionnaire, through the three selected dimensions: *quality*, *effort* and *acceptance*. We then present the results from correlation analysis before discussing further observations. Statistical significance was computed for each questions with a t-Test, two-sample assuming equal variances. We report strongly significant ($p<0.05$) but also somewhat significant ($p<0.1$) results because this is a lab-study, and the post-interviews supported these values.

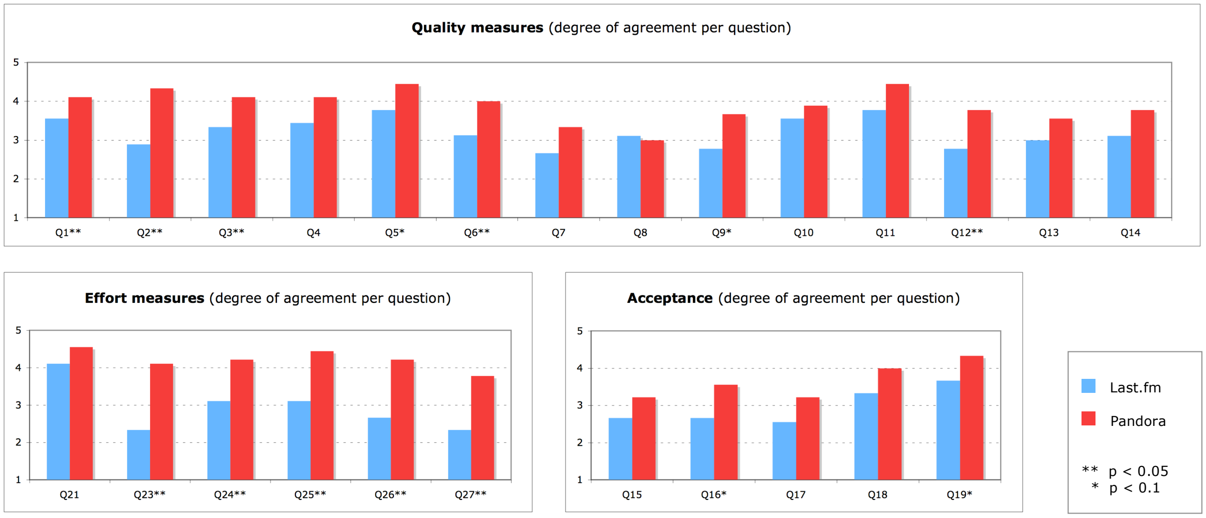


Figure 5. Agreement levels to postulates from final questionnaire

Perceived Usefulness - Quality

The results for the *quality* assessment are shown in the first graph of figure 5. The first surprising observation is that for each question where the difference between the two systems is statistically significant, the distinction is in favor of *Pandora*. Still, overall the answers are above 3, the average value of the agreement scale, which is a positive indicator that quality is a major dimension.

Possibly the most striking significant difference can be noted for Q2 where *Pandora* seems to be much better at providing songs that "suit a user's mood". The two averages have quite a large difference of 1.4. During the interviews, several people spoke about their mood to justify a certain interaction, positive or negative, with the tested system. Another notable difference can be seen with Q12 which checked with participants if they were "able to influence the quality of recommendations through their preference feedback". When asked in the post-study interview if they felt their feedback was taken into account, several *Last.fm* users mentioned that some artists they had banned were proposed again, sometimes even within the next five minutes. Furthermore, several of them simply said that they had not observed any difference from their feedback.

Another remarkable difference can be seen for Q9. Users perceived *Last.fm*'s recommendation technology as being less accurate than that of its counterpart. Although only somewhat significant, this is supported by post-study interviews where *Last.fm* users often reflected negatively on the accuracy during the post-study interviews, contrary to several *Pandora* subjects who were curious to know more about how the system was achieving such good results.

The differences for Q1 and Q6 were also found to be significant. *Pandora* participants thought that the recommended songs were more enjoyable and more satisfying. For both systems, users considered that songs were novel and above the middle value of the Likert scale, but the difference was not significant. We did not find this very surprising since playing songs randomly, with no personalization, can also guarantee novelty. The quality dimensions which could help explain the users' higher enjoyment and satisfaction with *Pandora* might come from Q3 and Q5. With Q3, users claimed that songs were better tailored to their taste, and in Q5 they stated that the recommend songs that were already known, were songs that they liked. For Q5, the difference was only somewhat significant.

Perceived Ease of Use - Effort

The results for the *effort* assessment are shown in the first graph of the second line in figure 5. These questions are also measured on the Likert scale, but the scale was reversed for Q21, Q23, Q26 and Q27. The general observation for this graph is that *Pandora*'s measures are very high, where as those from *Last.fm*, to the exception of Q21, are average, if not sub-average.

All questions have statistically significant differences, with the exception of Q21: users from both systems found that the registration process did not require too much effort, since both averages are above 4 out of 5 (reverse scale), with no significant difference. This is coherent with textual answers from Q20. The biggest difference comes from Q23: *Last.fm* users found that the initial time for the system to propose interesting music was too long contrary to those from *Pandora*. Their difference between the systems' averages is 1.8. Q22 textually questioned users about this dimension. *Last.fm* people report times between 15 minutes and 3 hours, with even two users saying that they had not yet received interesting music at the end of the experiment. On the contrary, *Pandora* users report times between 3 and 12 minutes, with the exception of one user who couldn't find her preferred style of music (oriental).

The users were also more in agreement that *Pandora* was easy to use as an "internet radio for listening to music" (Q24) and as a "recommender system for discovering music" (Q25),

Perceived Usefulness - Quality	
Q1	The songs recommended to me were enjoyable.
Q2	The songs recommended to me suited my mood.
Q3	The songs recommended to me were tailored to my taste.
Q4	The songs recommended to me were novel.
Q5	The recommended songs that I already knew are songs I like.
Q6	In general, I am satisfied with the songs recommended to me.
Q7	The recommended songs are as good as those I would receive from my friends.
Q8	Too many recommended songs were similar to each other.
Q9	The system's recommendation technology is accurate.
Q10	The system has enough music to propose as recommendations.
Q11	I like the fact that the system elicits preferencs from me.
Q12	I am able to influence the quality of recommendations through my preference feedback.
Q13	The system understands my musical taste and preferences.
Q14	I am able to determine how the system recommends music to me after using it.

Perceived Ease of use - Effort	
Q20	How long did it take you to register?
Q21	The registration process required too much effort. (*reverse*)
Q22	How long did it take for the system to initially propose the first few enjoyable songs?
Q23	The initial time it takes for the system to recommend interesting music is too long. (*reverse*)
Q24	The website was easy to use as an internet radio for listening to music.
Q25	The website was easy to use as a recommender system for discovering music.
Q26	The website offers too many features which are not relevant to music recommendations. (*reverse*)
Q27	There were too many navigable links which made the website confusing. (*reverse*)

Acceptance	
Q15	If a similar technology existed for recommending other things to me (books, movies), I would use it.
Q16	I would like to own the recommended songs.
Q17	I would purchase the recommended songs given the opportunity.
Q18	I found this website useful for listening to music that I like therefore I will use it again.
Q19	I found this website useful for discovering new music that I like therefore I will use it again.

Table 1. List of the Questions from the study

with a bigger difference for this second question. It is further supported by a difference in median values, unlike Q24. Finally it seems that contrary to *Pandora* users, *Last.fm* subjects were in agreement that "the website offered too many features which are not relevant to music recommendations" (Q26) and that "there were too many navigable links which made the website confusing". During the interviews, several mentioned being confused on the website and didn't know where and when to click or not. This observation was made independently of the users' level, normal or expert.

Acceptance

The results for the *acceptance* assessment are shown in the second graph in line two of figure 5. Of the three graphs, it is the one where the values are the lowest on average.

The differences for Q19 are only somewhat significant, indicating that users found *Pandora* a bit more useful for discovering music than *Last.fm* and were hence more inclined to use it again, but for Q18 the difference was not meaningful. The mean scores for both systems reveal that users find them useful for listening to and discovering music. The scores for the other questions are not as high. People only half agreed than "if similar technology existed for recommending other items (books, movies) then they would use it" (Q15), and similarly they were not really prepared to purchase the songs given the opportunity (Q17). We find these results interesting as they tend to show that people find these recommendation techniques useful (for listening or discovering) but are still fairly hesitant when it comes to going one step further, like buying.

Correlation Analysis

We performed a correlation analysis of results, considering all results from both systems together. We found numerous correlations; the three following subsections present the main relationships between the questions from the *effort* with the *quality* dimension (table 2), the *quality* (table 3) and the *effort* (table 4) dimensions with that of *acceptance*. For enhancing readability, only statistically significant results are reported in the tables.

Effort & quality

Because of the high number of correlated questions between *effort* and *quality*, and in order to keep the data readable, we chose to only report the correlations significant at the 0.01 level in table 2. The strongest link comes from the ease of use of the website as an internet radio (Q24) which strongly correlates with Q1,Q2, Q6 and Q9. The ease of use for discovering music is also highly linked with Q1 and Q6. The initial time a system takes to recommend good music (Q23) is correlated with enjoyability and satisfaction (a long initial time reduces enjoyability and satisfaction). Surprisingly the number of features (Q26) and navigable links (Q27) correlate inversely with having songs suited to a user's mood.

Correlation of effort with quality

	Q1	*Q2*	*Q6*	*Q9*
Q23	.600**	-	.782**	-
Q24	.692**	.595**	.678**	.632**
Q25	.680**	-	.678**	-
Q26	-	.609**	.608**	-
Q27	-	.616**	-	-

∗∗ Correlation is significant at the 0.01 level (2-tailed)

Table 2. Correlation: PEOU (effort) with PU (quality)

Quality & acceptance

The enjoyment of songs (Q1), and having them tailored to one's taste (Q3) are clearly the two factors that most influence *acceptance*. We found that Q1 was correlated with the wish to own (Q16), the PU for listening (Q18) and the PU for discovering (Q19). Q3 was found to be correlated with exactly the same dimensions. In parallel with enjoyment, satisfaction (Q6) was also highly correlated with PU for listening (Q18) and discovering (Q19), and Q11, the fact that a system elicits users' preferences is also correlated with Q18 and Q19. Finally, the perceived accuracy of the technology is important as it is the only quality element that correlates with the intention to purchase (Q17). We believe that all these results support the idea that the quality of recommendations is a key issue in the recommendations process that leads to user acceptance, and this in particular through generating a perceived usefulness.

Correlation of quality with acceptance					
	Q15	*Q16*	*Q17*	*Q18*	*Q19*
Q1	-	.470*	-	.555*	.679**
Q3	-	.499*	-	.576*	.593**
Q4	.549*	.568*	-	-	-
Q6	-	-	-	.485*	.635**
Q9	-	-	.513*	-	.489*
Q11	-	-	-	.612*	.507*

∗∗ Correlation is significant at the 0.01 level (2-tailed)
∗ Correlation is significant at the 0.05 level (2-tailed)

Table 3. Correlation: PU (quality) with acceptance

Effort & acceptance

Not all of the *effort* questions correlated with *acceptance*, but those that did correlated strongly. A short initial time for generating interesting recommendations (Q23) correlates with PU for listening and PU for discovering (Q18 & Q19). It is even more impressive how Q25, the ease of use of the website as a recommender system for discovering music, correlates with all acceptance measures with the exception of Q15. We believe these results support our hypothesis that effort is a key issue in acceptance.

Correlation of effort with acceptance					
	Q15	*Q16*	*Q17*	*Q18*	*Q19*
Q23	-	-	-	.486*	.518*
Q25	-	.582*	.500*	.768**	.721**

∗∗ Correlation is significant at the 0.01 level (2-tailed)
∗ Correlation is significant at the 0.05 level (2-tailed)

Table 4. Correlation: PEOU (effort) with acceptance

Discussion and Future Work

The results are very impressive as under the current setup of this study, they show an unanimous "win" for *Pandora* across all tested dimensions, although some where not statistically significant. The following discussion section takes a look at some reasons "why" *Last.fm* is outperformed and how this all relates to the TAM.

The TAM postulates that the PEOU influences the PU, and that both influence the behavioural intentions to use a system. The results strongly support this linkage as the correlation between effort (PEOU) and quality (PU) is highly favourable since the two direct assessments of "ease of use" (Q24 & Q25) are very strongly related to respectively four and two main quality questions, including the the two direct questions on "usefulness" (enjoyability and satisfaction). This result is in total accordance with the TAM in terms of the link between PEOU and PU. The next links in the TAM, PEOU and PU with behavioural intentions to use the system (acceptance), are also clearly supported. Four PU questions (Q1, Q3, Q6 & Q11) and two PEOU questions (Q23 & Q25) present strong correlations with the acceptance questions Q18 and Q19. We believe that these results clearly show that the TAM is an excellent model for music RS which captures in a simple way the core interaction dimensions in the acceptance process of such a system. What is surprising is how questions which directly asses the simple dimensions of this model, usefulness and ease of use, are systematically highly correlated, unlike indirect questions. This seems to indicate that although Davis et al.'s TAM is very basic and not recent, it still encapsulates the major and fundamental components leading to acceptance.

This having been said, and in accordance with points made in related work, there are some significant domain-specific elements which don't quite fit the model and where an extension to the TAM has to be considered. In this study, a miss-fit can be seen with Q9: the perceived accuracy of the underlying algorithm is the only value which correlates with the intention to purchase the recommended song, given the opportunity (Q17). This is interesting because for both systems, the score of the perceived accuracy of the recommendation technology is below that of the average quality question. Yet, questions such as Q1 or Q5 and the post-study interviews reveal that overall users were satisfied with either system, and dimensions like novelty show good correlation with acceptance questions. Based on these results it seems that in order to please users, the system only needs to have a "minimal" recommendation quality, which should of course take into account elements such as novelty (or diversity, as shown in [14, 9]). But in order to get users to go further in the acceptance process and actually buy songs, the system's recommendation accuracy seems crucial, whilst maintaining an easy to use system. We believe that this is an important result and will be explored in the future work.

Interestingly, *effort* results show that *Last.fm* users don't find the recommendations adapted to their mood (contrary to *Pandora* testers), and that the number of features (Q26) and navigable links (Q27) correlate inversely with having songs suited to a user's mood. This is quite surprising as one could easily assume that by providing the user with more tools to input his preferences, the system's recommendations should get more precise thus closer to his current mood. HCI has always had to find the balance between control and ease of use. In the case of *Last.fm* it seems that they have gone over the "tipping point" where small features are actually hampering the end-user's experience. *Last.fm* is clearly a successful website with more than ten million users. However, based on our results we believe that this does not primarily come from the recommender system which clearly poses some problems, but possibly more from the website's social features (which were not captured by this study). This issue will be addressed in future work.

Possibly the most significant contribution of this paper lies in the results for Q2: *Pandora* seems to be much better at providing songs that "suit a user's mood". In today's RS, the default mechanism for users to provide feedback is based on providing a kind of *score* either on a rating scale, or as with these music RS as a positive / negative score ("I like this song" / "I don't like this song"). However several studies in psychology, linked to music, have come up with different music classification schemes related to emotions. They propose new dimensions such as *arousal* and *valence* [7, 1]. The fact that the *mood* component is so prominent in our study supports the idea that the default positive / negative feedback process is not optimal and that there are other control dimensions which should be provided.

But beyond this potential future control mechanism, we believe that there are other reasons which explain why *Pandora* manages to "suit a user's mood" so well. When asked in the post-study interview if they felt that the their input was influencing the system, the users responded very differently: *Pandora* users answered "yes", whereas most *Last.fm* testers said "no". It seems that the responsiveness of algorithms is playing an important role here. Recommenders using collaborative filtering techniques are know to have computational issues with many users. These lead to profile updates being calculated offline, less frequently (as it is the case for *Last.fm*), on the contrary to content-based RS which can evolve immediately to users' input (*Pandora*). It therefore seems that for music recommender systems, the reactiveness of the system is a key component in the users' satisfaction. In the case of large systems, this would be a reason to favour a content-based approach.

CONCLUSIONS

Our study on user acceptance of recommender systems has very encouraging results. It is interesting since it compares two music recommenders and reveals key interaction features. Our results showed that users perceive *Pandora*'s recommendations as being more accurate, more suited to their mood, with songs more tailored to their taste, more enjoyable and more satisfying than those from *Last.fm*. In terms of effort, *Pandora* testers are equally satisfied with the initial time it takes for the system to recommend interesting music, and its ease of use for listening to and discovering music. In comparison, *Last.fm* users were less positive on several quality issues, and clearly unhappy with the initial time to reach good recommendations and the website's complexity, which proposed irrelevant features and was at times confusing.

The paper is one of the first to study acceptance issues in recommendation-seeking systems, for low involvement products. It reveals that F. Davis's initial Technology Acceptance Model can be successfully applied to such recommender systems. Although the model is quite simple, we show that it clearly suffice to capture the fundamental interaction mechanisms leading to acceptance.

This research points out two important results. First, the results show that overall user satisfaction in music recommender systems can be reached through several dimensions such as novelty. However the system's recommendation accuracy is the crucial component as it is the only one which correlates with the intention to buy the proposed songs. Second, while highlighting new control dimensions for music recommender systems, the study shows the necessity for low-involvement recommenders to be highly reactive. In this context, content-based recommenders appear to be most appropriate.

REFERENCES

1. ANDRÉ, S. K. . E. Composing affective music with a generate and sense approach. In *Proceedings of Flairs 2004 - Special Track on AI and Music* (2004), AAAI Press.
2. DAVIS, F. D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly 13*, 3 (September 1989), 319–340.
3. DAVIS, F. D., BAGOZZI, R. P., AND WARSHAW, P. R. User acceptance of computer technology: a comparison of two theoretical models. *Manage. Sci. 35*, 8 (August 1989), 982–1003.
4. HEAD, M. M., AND HASSANEIN, K. Building online trust through socially rich web interfaces. In *PST* (2004), pp. 15–22.
5. JONES, N., AND PU, P. User technology adoption issues in recommender systems. In *Networking and Electronic Commerce Research Conference* (October 2007).
6. KAMIS, A. A., AND STOHR, E. A. Parametric search engines: what makes them effective when shopping online for differentiated products? *Inf. Manage. 43*, 7 (2006), 904–918.
7. LANG, P. J. The emotion probe: Studies of motivation and attention. *American Psychologist 50*, 371-385 (1995).
8. MARIOS KOUFARIS, W. H.-S. Customer trust online: examining the role of the experience with the web-site. CIS Working Paper Series CIS-2002-05, Zicklin School of Business, Baruch College, New York, may 2002.
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., RIEDL, J., AND KONSTAN, J. A. Making recommendations better: an analytic model for human-recommender interaction. In *CHI '06 extended abstracts on Human factors in computing systems* (2006), ACM Press, pp. 1103–1108.
11. 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.
12. SCHAFER, J. B., KONSTAN, J. A., AND RIEDL, J. Recommender systems in e-commerce. In *ACM Conference on Electronic Commerce* (1999), pp. 158–166.
13. VAN DER HEIJDEN, H. Using the technology acceptance model to predict usage: Extensions and empirical test, 1986.
14. 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.