Eye-Tracking Product Recommenders' Usage

Sylvain Castagnos
EPFL - HCI Group
IC IIF Station 14
1015 Lausanne, Switzerland
sylvain.castagnos@epfl.ch

Nicolas Jones
EPFL - HCI Group
IC IIF Station 14
1015 Lausanne, Switzerland
nicolas.jones@epfl.ch

Pearl Pu
EPFL - HCI Group
IC IIF Station 14
1015 Lausanne, Switzerland
pearl.pu@epfl.ch

ABSTRACT

Recommender systems have emerged as an effective decision tool to help users more easily and quickly find products that they prefer, especially in e-commerce environments. However, few studies have tried to understand how this technology has influenced the way users search for products and make purchase decisions. Our current research aims at examining the impact of recommenders by understanding how recommendation tools integrate the classical economic schemes and how they modify product search patterns. We report our work in employing an eye tracking system and collecting users' interaction behaviors as they browsed and selected products to buy from an online product retail website offering over 3,500 items. This in-depth user study has enabled us to collect over 48,000 fixation data points and 7,720 areas of interest from eighteen users, each spending more than one hour on our site. Our study shows that while users still use traditional product search tools to examine alternatives, recommenders definitely provide users with new opportunities in their decision process. More specifically, users actively click and gaze at products recommended to them, up to 40% of the time. In addition, recommendation areas are highly attractive, drawing users to add 50% more items to their baskets as a traditional tool does. Observing that users consult the recommendation area more as they are close to the end of their search process, it seems that recommenders enhance users' decision confidence by satisfying their need for diversity. Based on these results, we derive several interaction design guidelines that can significantly improve users' satisfaction and perception of product recommenders.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*Human Information Processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/Methodology*

General Terms

Theory

RecSys2010, September 26–30, 2010, Barcelona, Spain.

Keywords

Time-dependent Model of Diversity, Impact and Evaluation of Recommenders in Practice, User Studies, Decision Theory, Usage Patterns and Attention

1. INTRODUCTION

Recent studies have shown that accuracy is not the only dimension to be taken into account when measuring the quality of recommender systems [10]. For example, diversity can be seen both as an obstacle and a complement to accuracy [17]. We here aim at understanding these needs for accuracy and diversity within the e-commerce environment by employing an eye-tracking system.

Most of the time, this type of research tends to evaluate how recommenders can adapt to users. Very few studies aim at understanding the reverse, whereby recommenders change the way users browse an online shop and how they capture the attention of users. In this paper, we seek to extend the purchase decision theory by analyzing users' behaviors as they interact with a website providing a multi-criteria filtering tool (MCF) and a recommender system (RS). The MCF search tool highlights the interactions through a classical interface that does not supply personalization, but helps users to reduce the number of displayed products from a set of constraints. The RS system aims at presenting relevant alternatives to a given product with several levels of diversity, thus assisting users in their choices. We relied on a critique-based recommender – called EPC [16] – which offers the advantage of eliciting users' feedback from a set of critiques within a session to improve accuracy. We set up an in-depth lab-study with an eye tracker and observe the behavior of eighteen users. Each of the users would spend up to one hour looking for a preferred product on a perfume e-commerce website.

This work represents a major step towards the formalization of sub-processes involving recommender systems within the purchase decision model. Following our previous works on the influence of the recommender on users' search and decision behavior over time [1, 2], we demonstrate that the increasing use of RS is explained by a need for diversity that leads to a higher confidence level and helps users reach a decision.

2. STATE OF THE ART

Several descriptive models seek to capture buying behaviors. Current literature reveals that there are six fundamental stages within the purchase decision making process [12] :

- First, users become aware of a new need. This realization can result from companies' prospecting campaigns or the recommendation of a new product from friends.
- Users will then determine from whom to buy, a process cal-

led merchant brokering. In online environments, consumers can use a price comparison website to determine where to buy their goods. In this case, everything that customers experience becomes an essential building block of a rapport between a buyer and a seller. Users are likely to look for website qualities that promote trust, ease of navigation, and strong relevance of items recommended to them [12].
- In the meantime, they evaluate product alternatives in order to make the final choice. At this stage, called *product brokering*, interactions help recommender systems to understand their needs and present personalized options to them.
- Stages 4 and 5 consist of negotiation and purchasing. The seller has to provide security and confidence in order to close the sale.
- Finally, users' satisfaction in relation to the overall buying experience can be measured in a sixth stage, if post-purchase product service is involved.

In this paper, we will focus on the *product brokering* stage where consumers evaluate product alternatives in order to make the final choice. At this stage, tracking users' interactions can help a recommender system understand their needs and present personalized options to them. According to Haubl *et al*., the product brokering stage can be divided into two steps [4]. During the first step, the active user identifies a subset of products to compare. During the second step, the different features and details of these products are compared in order to make a decision. Haubl also proved that the use of a recommender system leads to a reduction in the number of alternatives considered seriously for purchase [5], and that a recommendation agent increases the number of non-dominated alternatives – *i.e*. not objectively inferior to any alternative [15] – in the set of alternatives seriously considered for purchase. Based on such research, it is apparent that recommenders prove to be a useful tool to users, assuming that they provide items relevant to users' needs. However, the goal of personalization is not only to provide the right item to the right person, but also *right away* and at the right *time*. The time constraint has long been overlooked by researchers. In this paper, we aim to analyze both the impact of recommenders over time at the product brokering stage, thus extending the findings of [4]), and the factors that influence the users. The research questions with regards to the setup of a recommender focus less on *what* to suggest, but rather *when* and *why*.

The work outlined in [7] represents a pioneering effort to study the impact of personalization at different decision making stages. Through an experiment involving a ringtone personalization service, they highlight the decreasing probability of a tailored item to be selected at later stages of decision making. Nevertheless, the absence of selection does not mean that the recommender system does not play a role in the decision process ; merely looking at a recommendation can affect the user's ultimate decision. In [1], we showed that the influence of a recommender in comparison with a MCF tool constantly increases as time goes on. We demonstrated that the influence of RS is in fact maximal when the active user is close to making a decision and adding a product to the basket. This sheds light on the *when* to recommend question.

The conclusions in [9] underline an inappropriate combination between accuracy and diversity at later decision stages. This brings us back to the question of *why* to suggest items. Ho *et al*.[8] showed the influence of the need for cognition [1] and the size of the recommendation set on decision making. In this paper, we will investigate the need for diversity in order to reach a decision. Diversity is an ongoing topic of discussion in the arena of recommender research, although it has been explored relatively less than other dimensions of the recommendation process. It is agreed that diversity has a role to play in making good recommendations and is thus a sought-after property, but *why* and *how* to use it is a disputed topic. The first major paper on the matter was likely [13] where, in the context of conversational recommender systems, they showed that introducing diversity significantly enhances the efficiency of recommendations. In [18], Ziegler and McNee introduce a method for designing and diversifying personalized recommendation lists, thus decreasing average accuracy but increasing user satisfaction. Even very recent user studies such as [11] point out how necessary diversity can be. Thus, we chose users' need for diversity as our dimension for investigating the *why*.

1. The need for cognition is a personality variable reflecting the extent to which people engage in and enjoy effortful cognitive activities.

3. EXPERIMENT SETUP

3.1 The Material

The experiment consisted of an in-depth real-user evaluation with an eye tracker on a perfume e-commerce website. The eye tracker used in our experiment was a `Tobii 1750`. This device consists of a computer screen with a camera installed on the top. Except for a short calibration phase, the setup allows users to look at the screen in a natural way without the need for a head mount.

The perfume database consists of more than 3,500 perfumes, which contains all popular brands and perfumes that are available in regular perfume shops. Information on popularity, quantity, brand, price, and other product characteristics of each perfume was carefully selected and included. The perfume domain was chosen as it is a slightly above-norm field in terms of complexity. Had a very common domain been selected, users would have felt less engaged in their interactions, possibly resulting in some "shortcut" behaviours. Furthermore, by having a less conventional domain, users are forced to rely more on the tools proposed, helping us to evaluate the efficiency of the different parts tested. Finally, most users are not perfume experts and have stable preferences with regards to these public taste products.

Wanting to analyze the behaviors of users in a realistic situation, the design of the website was chosen to reproduce the template of a classical e-commerce application such as Amazon.com. Thus, the site contained two main windows, as shown in Figure 1.

The first window, called the *search page*, was divided into two parts. At the top of the page, a multi-criteria search tool (MCF) was conceived that included brands, price ranges, quantity ranges and types of perfume (*eau de parfum*, *eau de toilette*, *aftershave*). Below this, a double-column, lexicographically ordered item-list of perfumes was available. This part displayed the perfumes respecting the selected criteria of the upper multi-criteria search tool. In this double choice-list, each perfume was laid out with a picture of the bottle, its exact name, brand, price and quantity. The first two lines of its description were also shown. In addition, a classical re-ordering tool allowed users to sort the list of results by brand, price (low to high, or high to low), and popularity. The search page also included a possibility to set the number of results displayed per page and the currency. By default, the results were displayed in US dollars and sorted by popularity, with sixteen presented at a time.

The second window, hereafter the *detail page*, showed the information about any perfume. In addition to presenting the same information as in the list view (see above), specific data was provided here including : a full description, a big-sized picture, a best-selling rate, average user ratings, gender, source website, and the possibility to rate. The page had an "Add to shopping cart" button. To the right of this detail, a column of recommendations was si-

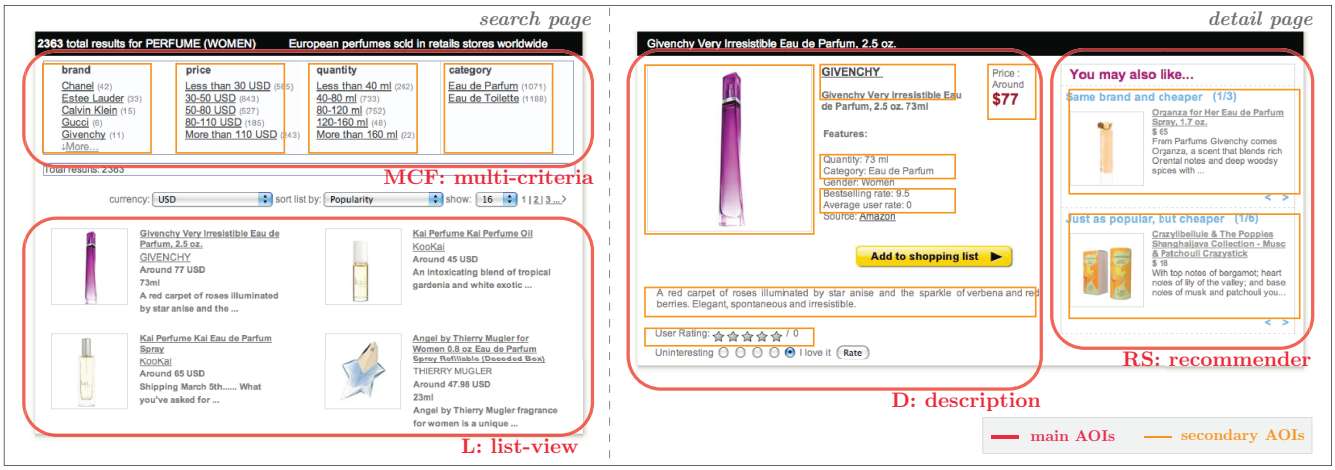

Figure 1: Snapshots of the main interfaces (inc. AOIs) : the search page (a), and the detail page (b)

multaneously proposed in five classified boxes, all labeled either : "more popular and cheaper", "more popular but more expensive", "same brand and cheaper", "same brand but more expensive", "just as popular and cheaper", "same price range and just as popular" or "people who like this also like". Although seven types of recommendations were available, we chose to display only five at any given moment in random order, in order to reduce users' habit of their screen positions. Each box contained up to six items, which were horizontally scrollable without reloading the page.

The first six recommendation categories were generated from editorial picked critiques (EPC) [16], adapted from the organized critiquing method [3] with users' needs for popularity and editorial information. This kind of algorithm is well-adapted to our product domain, since it allows explanations for recommendations to users in a context where they cannot smell perfumes and are consequently more difficult to convince. Since we know that the quality of recommendations has a great impact on click-through rate, we chose to rely on the EPC algorithm which has been shown to be preferred 2.42 times more than general critique-based recommenders [16]. These recommendation categories are related to features of the products. Compound critiques are automatically generated by altering values of some variables – keeping others constant – and computing trade-off benefits compared to the current product. The seventh category "people who like this also like" relied on a standard collaborative filtering algorithm. Our recommender system (RS) provided categories with different levels of diversity, as explained in the following subsection.

3.2 Diversity Metrics

Diversity can be defined in many ways. In this work, we consider that diversity is a measure which inversely relies on the similarity between products [18]. The more two items are similar, the less diversity there is between them. Drawing our inspiration from [17], we compute the similarity between two products p_1 and p_2 by using a weighted mean of the similarities between p_1 and p_2 for each of the five attributes that characterize a perfume (brand, price, quantity, category, and popularity), as shown in Equation 1.

$$Sim(p_1, p_2) = \frac{\sum_{i=1..5} w_i \ * \ sim_{attribute=i}(p_1, p_2)}{\sum_{i=1..5} w_i} \quad (1)$$

We then computed the average intra-list similarity (ILS) of each recommendation category C, by adapting the measure defined in [18]

Table 1: Average Intra-List Similarity for six recommendations (ILS), Average Similarity between a recommendation and the perfume of the current detail page (Sim), and Relative Diversity of a recommendation relatively to the current perfume (RD)

Category	ILS	Sim	RD
More popular and cheaper	0.6	0.4	**0.6**
More popular but more expensive	0.6	0.4	**0.6**
Same brand and cheaper	0.6	0.6	0.4
Same brand but more expensive	0.6	0.6	0.4
Just as popular and cheaper	0.6	0.6	0.4
Same price range, just as popular	**0.8**	0.8	0.2
People who like this also like	**0.4**	0.4	**0.6**

(see Equation 2). If card(C)=n, then :

$$ILS(C) = \frac{\sum_{p_i \in C, i=1..n-1} \sum_{p_j \in C, j=i+1..n} Sim(p_i, p_j)}{\frac{n \ * \ (n-1)}{2}} \quad (2)$$

Notice that a higher ILS score denotes lower diversity. We also computed the relative diversity RD of a perfume p_i compared to a set of n perfumes P, using Equation 3 [13].

$$RD(p_i, P) = \frac{\sum_{j=1..n} (1 - Sim(p_i, P_j))}{n} \quad (3)$$

The set of perfumes P represents products that have caught the attention of the active user on a given page. The relative diversity consequently allows us to measure the added value of considering a new perfume, relative to the sequence of past consultations.

Table 1 summarizes similarity and diversity scores of the seven recommendation categories. We calculated the average intra-list similarities (ILS) by considering that every category was composed of six recommendations on the detail page. This is an pessimistic estimate, since it corresponds to the maximal number of recommendations for a category ; with fewer items in a category, it is more diverse. Table 1 also displays the average similarity (Sim) and relative diversity (RD) between a recommendation and its related product. In the case where we consider the relative diversity between only two products, we can say that $RD = (1 \ - Sim)$.

The categories that provide the biggest diversity are "more popular and cheaper", "more popular but more expensive", and "people

Table 2: Average Relative Diversity between two perfumes coming from MCF tool

Number of selected criteria	0	1	2	3	4
Sort criterion $\in$ MCF selection	-	0.8	0.6	0.4	0.2
Sort criterion $\notin$ MCF selection	0.8	0.6	0.4	0.2	0

who like this also like". On the contrary, the category "same price range and just as popular" supplies very poor diversity, but higher similarity.

At last, we computed the average relative diversity between two products coming from the multi-criteria search tool (see Table 2). This value is dependent on the number of selected criteria and the way products are sorted in the list view (sort criterion linked to selected MCF criteria or not).

The use of these metrics supposes that we are able to measure the interest of users for the different products proposed through the interface. This has been done through the use of action logs and gaze data.

3.3 Experiment Procedure

The study was designed as an in-depth one-hour lab study. At all times, participants could ask questions and obtain answers from the available assistant conducting the study. The general online evaluation procedure consisted of the following steps :

Step 1. The participant is welcomed by the assistant. He is briefly introduced to the topic of the experiment, which was described as a user study in perfume preferences. He is informed that the perfume e-commerce website he will test contains over 3,500 most commonly used and sold perfumes in the world. He is also told about the incentive for completing the study (see below).

Step 2. The user is asked a detailed set of background questions (age, sex, etc.).

Step 3. The eye tracker is calibrated to the user's eyesight. The experiment begins and the tracking session is launched by the assistant, who encourages the user to explore the system before fully launching into the first task.

Step 4. The user then has two separate tasks to complete in two sessions (`Session 1` and `Session 2`). One goal is to select up to three perfumes that he has never heard of or used before, but that he would be prepared to buy for himself. He is asked to put them in the basket, and informed that he may select more than three and delete some at the end. In the rest of this paper, this recording will be called `Task S` (Self). The other goal consists in searching for one perfume he would like to offer to someone, preferably of the opposite sex, to reduce potential bias of product habituation. This will be called `Task G` (Gift). In order to reduce another bias linked to fatigue, we alternate the order of sessions. Half of the users complete the `Task G` in `Session 1`, before `Task S` in `Session 2`. The others start with `Task S` and end with `Task G`.

Step 5. To conclude the study, fourteen preference questions are asked in order to explicitly assess users' overall perceptions of the system after the experiment on a five-point Likert scale (cf. Table 3). Ratings vary from -2 (strongly disagree) to $+2$ (strongly agree), where 0 is neutral. This allows us to match explicit and implicit data to confirm our hypothesis. The questions were asked in random order to eliminate ordering bias.

3.4 Participants' Background

The user study was carried out over a period of three weeks. Immediate incentives, chocolate or wine, were offered directly after the study. More importantly, users who had completed the study took part in a draw for a CHF 100.- voucher to buy one of the perfumes they had added to the basket. By proposing this high value incentive, we ensured that users behaved truthfully throughout the selection process. A total of eighteen volunteers were recruited as participants. They were from three different continents, with different professions (student, worker, Ph.D. student) and educational backgrounds (high school, graduate school).

All users had strong web experience, although online shopping experience remained limited to classical items such as books, music, travel, and electronic items.

The second part of the background questions surveyed users' predisposition towards perfumes. Six participants said they bought perfumes about once a year, nine a few times a year and one nearly monthly. When questioned about how they discovered new perfumes, 61% said that they preferred to test perfumes alone in a shop. And finally, six of the eighteen participants considered themselves to be experienced in perfumes. In the rest of the paper, we will call them "perfume experts".

Table 3: Post-Stage Assessment Questionnaire

ID	Statement on recommended items
P1.	The recommended items are attractive.
P2.	The recommended items are educational.
P3.	The recommended items appeared to be a good deal.
P4.	The recommended items appeared to be marketing material.
P5.	The recommended items influenced my selection.
P6.	The recommended items will influence my future selection.
P7.	The names of the categories are useful and adequate.
P8.	I am satisfied with the overall quality of the interface.
P9.	I found the interface easy to use.
P10.	I would buy the perfumes recommended to me, given the opportunity.
P11.	If this were a real website, I would use it in the future to find perfumes.
P12.	I believe that the recommender algorithm is efficient.
P13.	The recommended perfumes were diverse.
P14.	The recommended perfumes were novel.

3.5 Definitions

This subsection introduces all the definitions required to measure the impact of the recommender system.

DEFINITION D1.

We define a recommendation category as **dominating** *if :*

$$\%\ of\ usage\ of\ a\ category > \frac{100\%}{No.\ of\ categories}$$

DEFINITION D2.

We define a recommendation category as **influential** *if a basket product comes directly from a recommendation of the considered category.*

DEFINITION D3.

We define the **gaze rate** *as the percentage of consulted detail pages where the active user examined the recommendation area.*

DEFINITION D4.

We define the **click rate** *as the percentage of consulted detail pages where the active user clicked on a recommendation.*

DEFINITION D5.

We define the **exploration rate** *as the percentage of consulted detail pages where the active user browsed the different recommendations available in at least one category (by clicking on the arrows "previous" or "next" of one or more categories).*

4. GOALS OF THE EXPERIMENT

In [1], we showed that the recommender has an impact on two general aspects of consumer decision making in an online shopping environment : (1) choice strategies which can be thought as sequences of operations for searching through the decision problem space [15], and (2) consideration sets conceptualized as alternatives that consumers consider seriously for purchase [6]. We also proved in [1] that the influence of the recommender system continuously increases over time at the product brokering stage.

The goal here is to go into further detail to understand the exact role of the recommender within the decision process. In particular, we aim at understanding what aspects of the recommender help make the purchase decision simpler. We believe that when users are about to make a decision (about a purchase), recommender systems help users to increase their confidence, by fulfilling their need for diversity (among similar items).

Extending our previous works [1] proving that the influence of the recommender is maximal when users are close to making a decision, we here hypothesize that basket products will more often come from the interactions with the recommender than from pure interactions with the multi-criteria search tool. This would strongly support that the recommender increases user confidence, *i.e.* the ability to choose an item among valuable alternatives. Moreover, we intent to show that the influence of the recommender is explained by the users' need for diversity. We consequently expect users to prefer the recommendation categories providing the greatest diversity.

5. RESULTS AND DISCUSSION

Using the eye tracking system, we first aimed at measuring how users' interest for the different parts of the website evolve as time goes on. Apart from the usual gaze plots and heat maps which can be collected with an eye tracker (see Figure 3), we decided to rely on a large palette of *Areas Of Interest* (AOI) as shown in Figure 1.

Averaging 1,350 fixations per user and per session, we recorded 48,891 fixation points throughout the study. We defined 7,720 AOIs, sorted into 593 different pages. These were two different kinds of pages as explained above : the search pages and the detail pages. The statistics of usages between `Session 1` and `Session 2` are very similar (3.7% of difference as regards the number of viewed pages). Because the experiment was quite long, we checked this potential difference and alternated `Task S` and `Task G` to dismiss fatigue as a factor influencing the users' behaviors.

We then computed the total fixation durations for each user on the different AOIs over time `t`. We paid particular attention to durations for two variables : the multi-criteria box `MCF` and the recommender system `RS`.[2] Usages of the MCF and RS over time are made explicit in Figure 2. We summed the cumulative fixation durations of these two AOIs. We noticed that the use of RS increases much faster than the use of MCF as time goes on. RS is used 35% more than MCF at the end of `Session 1`. This is significant at 0.99 level ($p = 0.005$). The difference is even bigger in `Session 2`, where the RS remains stable and is used 88% more than MCF ($p = 0.076$). This confirms the important role of RS in the purchase decision making process.

As a continuation of these observations, we examined the impact

2. The usage of additional AOIs such as the list view and the product description are discussed in [2].

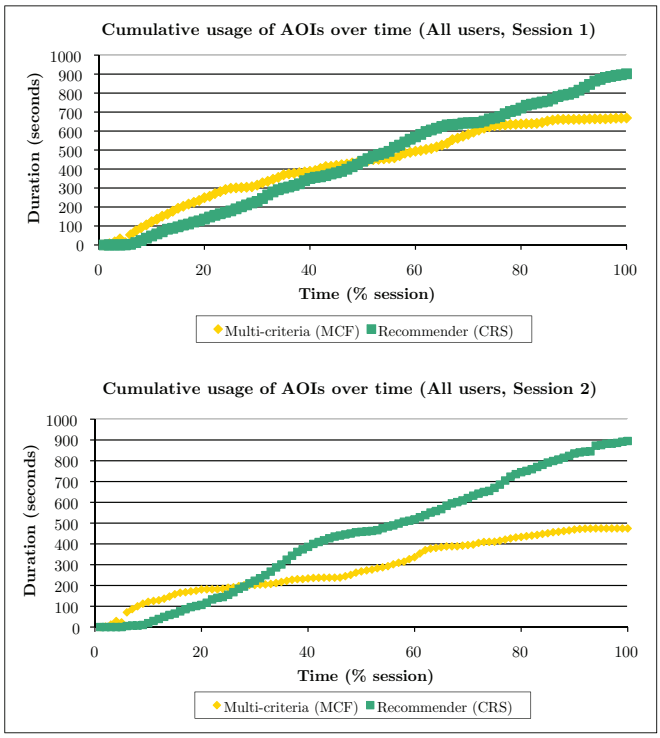

Figure 2: Usage of RS and MCF over time for the overall set of users

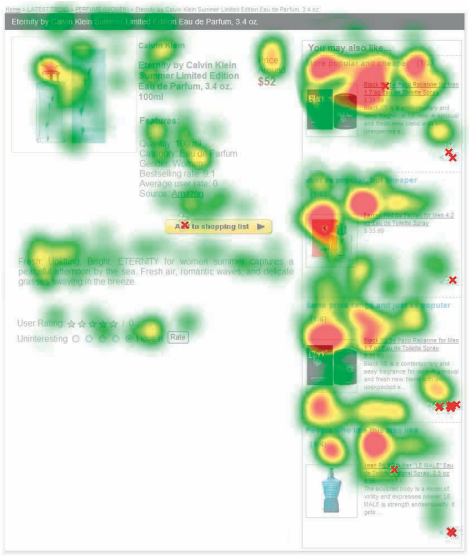

Figure 3: Example of heat map : colors vary according to where users looked most, red crosses show clicks.

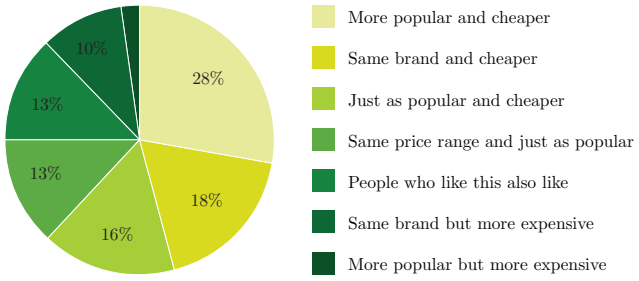

Figure 4: Proportion of time spent looking at each category (all users, both sessions)

Table 4: Average effects of recommendation categories

	Number of dominating cat.	No. of cat. looked at	No. of rec. categories used	No. of influential categories
Task S	2.74	5.89	2.74	1.37
Task G	1.95	4.42	0.84	0.26
Both	2.34	5.16	1.79	0.82

of different categories in our RS. We first aimed at determining if eye movements revealed some dominating and influential recommendation categories within the recommender.

Heat maps allowed us to measure the impact of each of the seven recommendation categories on the overall set of users. We computed the number of times each user looked at a recommendation category. Figure 4 shows that each category has a significant importance when we merge all users' data, since most of them represents at least 10% of gaze interactions.[3] However, customers seem more attracted to items labeled as "more popular" and products from "people who like this also like". To avoid an overgeneralization based on global evidence, we went into further detail by examining the arithmetic means of recommendation categories' usages for all users (see Table 4). We compared the number of dominating recommendations (D1) with the number of recommendation categories looked at, the number of recommendation categories used (when users click on items of these categories), and the number of influential recommendation categories (D2).

On average, we can see that each user paid attention to five categories during the two sessions. However, only two of them can be considered dominating, and only one category influenced him. As expected, all the categories of RS are thus useful during the product brokering stage, but users filter the information to focus on categories in accordance with their individual needs and expectations. To understand how and why they individually favor a subset of categories, we analyzed the action logs. Clicks confirmed the preferences of users for categories "more popular and cheaper" and "people who like this also like", despite the fact that the most selected MCF criterion is the brand. Table 5 synthesizes the average number of clicks on MCF criteria and recommendation categories. Table 5 further points out users' disinterest in the categories "same price range and just as popular" and "just as popular and cheaper". Taking account the diversity scores reported earlier (see Tables 1 and 2), it appears that the increasing influence of RS is characterized by an attraction for categories providing the highest levels of diversity, supporting the idea that users are open-minded to diversity. The most popular categories of RS offer the same level of diversity as the list-view with one selected MCF criterion, but with a higher

3. Note that there were fewer items that fitted into "more popular but more expensive", which explains the low percentage.

Table 5: Average numbers of clicks (all users)

		Task S	Task G	Both tasks
MCF	Brand	**2.1**	**1.3**	**1.7**
	Price	0.7	0.4	0.6
	Quantity	0.4	0.1	0.3
	Category	1.2	0.6	0.9
RS	More popular and cheaper	**0.9**	**0.4**	**0.7**
	More popular, more expensive	0.2	0.0	0.1
	Same brand and cheaper	0.7	0.1	0.4
	Same brand, more expensive	0.3	0.2	0.3
	Just as popular and cheaper	0.3	0.1	**0.2**
	Same price range, just as popular	0.5	0.1	**0.3**
	People who like this also like	**0.9**	**0.3**	**0.6**

Table 6: Number of basket products : RS vs. MCF

	Session 1	Session 2	Both
Added after MCF selection without RS	**11**	**17**	**28**
Added after having been influenced by RS	18	14	32
Added after interacting with RS	5	5	10
Total number coming from RS (influence + interaction)	**23**	**19**	**42**

accuracy. Such evidence strongly supports the idea that users' need for diversity led them to use the recommender, rather than the MCF tool.

Finally, in order to reach our experiment goals, according to which this provided diversity increases user confidence, we evaluated the proportion of basket products coming from RS, rather than from MCF (see Table 6). At the first encounter with the system (Session 1), users' reliance on the recommender agent seems the strongest. They consulted it more frequently and twice as many basket items came from RS than from its MCF counterpart. This is significant at 0.95 level according to Student's T-test (p=0.049). After users learned more about the domain knowledge, their reliance on both agents becomes somewhat comparable. Among the 70 products globally added to the basket, 28 perfumes only came from the MCF tool without any interaction with RS (40%). 60% of basket products consequently came from the RS : 32 products were added just after a click on a recommendation (influential category), 10 perfumes were added following one or several comparisons with some recommendations (by going on detail pages of recommendations and then going back to the previous page). To summarize, this means that users more often reach a sufficient level of confidence when the considered product comes from RS, rather than MCF. And users consult the RS because it provides diversity, thus helping to increase user confidence.

We cross-checked with the responses of the post-study assessment questionnaire summarized in Figure 5. In order to measure and maximize the veracity of decisions to add products to the basket, we asked users if they would buy the chosen perfumes given the opportunity – keeping in mind that one participant was going to win a CHF 100.- voucher to buy one the perfumes added to their basket – or at least go in a perfume shop to smell them and learn more about them. 56% of users agreed to buy given the opportunity ; 17% were not sure, but agreed to smell them before making a final decision. The answers of the assessment questionnaire (see Figure 5) showed that both standard and expert users found the recommender attractive (P1), educational (P2), useful (P7), easy to use (P9) and efficient (P13). However, standard users seemed much more influenced by RS than experts (P5 and P6). Standard users strongly expressed their satisfaction with regards to diversity (P14). The correlation analysis revealed that recommendation diversity (P14) is strongly correlated to intention to buy (P10) and to influence of selection (P5 and P6). These correlations are strong and statisti-

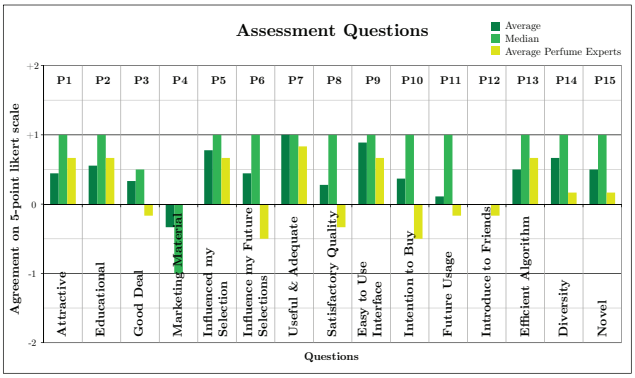

Figure 5: Answers of the Assessment Questionnaire

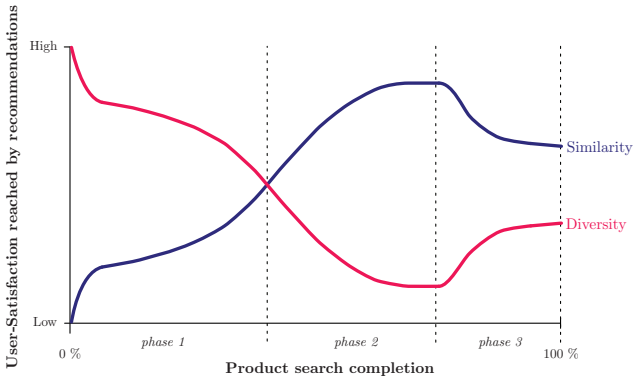

Figure 6: Time-dependent model of satisfaction : a dynamic compromise between accuracy and diversity in recommender systems

cally significant with respectively : $r = 0.547$ ($p = 0.015$) and $r = 0.579$ ($p = 0.009$). Finally, diversity is correlated with satisfaction ($r = 0.466$, $p = 0.044$) and ease of use ($r = 0.487$, $p = 0.034$). At last, we used a factor analysis to uncover the latent structure of the assessment questionnaire's variables. The cluster composed of the factors "diversity" (0.777), "intention to buy" (0.872) and "future influence" (0.704) explains for the most part the variability of answers. The reliability of this result is assessed by the Cronbach'alpha (0.756), with a significance of the Bartlett' Test of Sphericity equal to 0.001. The perceptions of users are consequently in accordance with the ideas proposed in this paper.

6. GUIDELINES

In this section, we aim to put forward the lessons learned from this study thanks to a series of design guidelines. Throughout this paper, we outlined how users progressively use the recommender system, and how as they get closer to their desired item, they need to explore recommended alternatives. We propose the following guideline :

Guideline 1 Consider providing accurate recommendations, *i.e.* close to users' concerns and expectations. This will encourage them to turn towards intelligent agents (at the expense of classical search tools such as MCF), thus improving their experience.

Consider providing recommendations which are accurate with regards to users' concerns and expectations. This will encourage them to turn towards intelligent agents (at the expense of classical search tools such as MCF), thus improving their experience.

In literature, it is admitted that a recommender system should maximize its accuracy with respect to users' preferences. In doing so, many approaches rely on similarity metrics [14]. Yet, in this paper we highlighted that accuracy is not the only criterion to take into account to maximize satisfaction. The need for diversity among recommendations also plays a central role in the decision process. A popular perfume like "Chanel n° 5" will exist in many different quantities and prices, meaning that similar recommendations will actually be the same perfume : this would not provide a good user experience. A more diverse set of recommendations would consist in proposing different perfumes from different brands for example. Each recommendation would have a high similarity with the user's profile, but a lower similarity with other recommendations. Consequently, increasing recommendation diversity reduces similarity between recommendations without necessarily lowering the accuracy. Smyth *et al.* have proven that recommendations' similarity and diversity are opposite dimensions [17]. The challenge while designing a recommender system, especially for entertainment products, is often to make such delicate tradeoffs between opposing requirements. We extend this conclusion by claiming that user satisfaction is a compromise between accuracy and diversity, and above all that it evolves over time. Said otherwise, users' needs change throughout the time of a decision process.

At first, the system knows little of nothing about the user's preferences. In accordance with Häubl's first choice strategy (discovery of adapted search criteria and valuable alternatives), we propose to maximize diversity as a first phase. In a second phase, users need accurate recommendations in order to get products as close as possible to their preferences or to the alternatives they had in mind. As suggestions become more accurate, the similarity should increase to approach the users' ideal. Then, they work towards establishing that this is a confident choice. In this third phase, if all the recommendations are very accurate, users will not be able to distinguish them, which makes decisions harder (despite the fact that all recommendations are adapted to users' expectations). By increasing the degree of diversity, we allow users to confirm their choices, or to extend their exploration towards items a little bit different from the current product.

Guideline 2 Consider favoring an approach where similarity between recommendations progressively increases during the first two phases, and where diversity is re-introduced in the third phase (when decision is close to being made).

Guideline 3 Consider including diversity while preserving a reasonable similarity between recommendations. This approach can significantly enhance users' confidence in the items they have selected. Our experiment shows that the diversity level should be between 40% and 60% (cf. Tables 1 and 5).

The figure 6 summarizes the time-dependent model of satisfaction that we propose as a lesson from our investigations. This will have strong repercussions for both researchers and practitioners, since efforts should be made to conceive new evaluation metrics and aggregation functions that deal with time and need for diversity in addition to similarity with users' profiles when computing recommendations.

At last, we showed that the interactions with recommender systems largely exceed clicks. During our experiment, we respectively got an average click rate and exploration click rate of 40.01% and

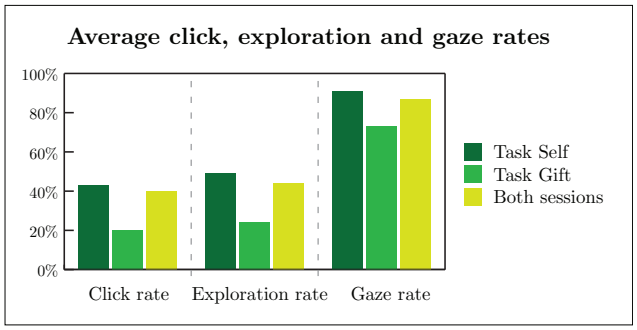

Figure 7: Modes of interaction with recommender systems

44.37% for both sessions (cf. figure 7). The gaze rate was much higher with an average of 87.20%.

User attention is consequently not restricted to clicked items. The simple fact of taking a look on the recommender system plays a role in users' cognitive processes and decision making process. We consequently propose the following guideline :

Guideline 4 Catching feedback in real time from implicit criteria such as eye movements (with an eye tracking system in active mode, when possible) can significantly improve the understanding of users' needs and behaviors.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we examined the impact of a perfume recommender system over the entire product brokering stage and showed how it changes customers' purchase decision. We looked very precisely at behaviors of eighteen users, each spending more than one hour on our site. We were able to track users' actions and collect almost 49,000 fixation points with an eye tracking system in this in-depth study. We then paid attention to how the influence of recommender systems integrates into the purchase decision making model. The analysis of our results, cross-checked with the users' assessment questions, leads to two major conclusions. First, users rely on the recommender to enhance their confidence in the purchase decision. Second, our work demonstrates that the accuracy of recommendations is not the only criterion for the success of a recommender system. We outlined users' need for diversity when making a decision. This not only supports the theory, according to which it is necessary to find a good compromise between accuracy and diversity in order to increase quality of recommendations, but also provides surprising guidelines about when this diversity is needed.

This study constitutes a major step towards the understanding and formalization of users' interaction behaviors with a recommender. We first aim at reproducing this experiment in different domains with different levels of diversity to fully validate our satisfaction model. We also plan to do a within-subject study in order to precisely quantify expected and perceived diversity levels at each phase of the product brokering. We will then use these results to adapt recommendations to both users' preferences and need for diversity.

8. REFERENCES

[1] S. Castagnos, N. Jones, and P. Pu. Recommenders' influence on buyers' decision process. In *Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys 2009)*, pages 361–364, New-York, USA, October 2009.

[2] S. Castagnos and P. Pu. Consumer decision patterns through eye gaze analysis. In *Workshop on Eye Gaze in Intelligent Human Machine Interaction (IUI 2010)*, Hong Kong, China, February 2010.

[3] L. Chen and P. Pu. Preference-based organization interfaces : Aiding user critiques in recommender systems. In *In Proceedings of International Conference on User Modeling (UM'07)*, pages 77–86, Corfu, Greece, June 2007.

[4] G. Haubl and K. Murray. Preference construction and persistence in digital marketplaces : The role of electronic recommendation agents. *Journal of Consumer Psychology*, 13(1) :75–91, 2003.

[5] G. Haubl and V. Trifts. Consumer decision making in online shopping environments : The effects of interactive decision aids. *Marketing Science*, 19(1) :4–21, 2000.

[6] J. Hauser and B. Wernerfelt. An evaluation cost model of consideration sets. *Journal of Consumer Research*, 16 :393–408, March 1990.

[7] S. Ho. Web personalization and its effects on users' information processing and decision making. PhD Thesis of the Hong Kong University of Science and Technology, 2004.

[8] S. Ho, D. Bodoff, and K. Tam. Timing of adaptive web personalization and its effects on online consumer behavior. *Information Systems Research*, 2009 (forthcoming).

[9] S. Ho and K. Tam. An empirical examination of the effects of web personalization at different consumer decision-making stages. *International Journal of Human-Computer Interaction*, 19(1) :95–112, 2005.

[10] N. Jones. *User Perceived Qualities and Acceptance of Recommender Systems : The Role of Diversity*. PhD thesis, Swiss Federal Institute of Technology (EPFL), 2010.

[11] N. Jones, P. Pu, and L. Chen, editors. *How Users Perceive and Appraise Personalized Recommendations*. Springer, 2009.

[12] P. Maes, R. Guttman, A. Moukas, and R. Moukas. Agents that buy and sell : Transforming commerce as we know it. *Communications of the ACM*, 42 :81–91, 1999.

[13] L. McGinty and B. Smyth. On the role of diversity in conversational recommender systems. In *Proceedings of the International Conference on Case-Based Reasoning Research and Development (ICCBR)*, pages 276–290, 2003.

[14] S. McNee, J. Riedl, and J. Konstan. Being accurate is not enough : how accuracy metrics have hurt recommender systems. In *CHI '06 : CHI '06 extended abstracts on Human factors in computing systems*, pages 1097–1101, New York, NY, USA, 2006. ACM.

[15] J. Payne, J. Bettman, and E. Johnson. *The Adaptive Decision Maker*. Cambridge University Press, 1993.

[16] P. Pu, M. Zhou, and S. Castagnos. Critiquing recommenders for public taste products. In *Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys 2009)*, pages 249–252, New-York, USA, October 2009.

[17] B. Smyth and P. McClave. Similarity vs. diversity. In *Proceedings of the 4th International Conference on Case-Based Reasoning*, pages 347–361, Vancouver, Canada, 2001.

[18] C.-N. Ziegler, S. McNee, J. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web (WWW 2005)*, pages 22–32.