The Effect of Context-Aware Recommendations on Customer Purchasing Behavior and Trust

Michele Gorgoglione
Polytechnic of Bari, Italy
Viale Japigia 182, 70126
m.gorgoglione@poliba.it

Umberto Panniello
Polytechnic of Bari, Italy
Viale Japigia 182, 70126
u.panniello@poliba.it

Alexander Tuzhilin
Stern School of Business, USA
44 West Fourth Street, NY 10012
atuzhili@stern.nyu.edu

ABSTRACT
Despite the growing popularity of Context-Aware Recommender Systems (CARSs), only limited work has been done on how contextual recommendations affect the behavior of customers in real-life settings. In this paper, we study the effects of contextual recommendations on the purchasing behavior of customers and their trust in the provided recommendations. In particular, we did *live* controlled experiments with real customers of a major commercial Italian retailer in which we compared the customers' purchasing behavior and measured their trust in the provided recommendations across the contextual, content-based and random recommendations. As a part of this study, we have investigated the role of accuracy and diversity of recommendations on customers' behavior and their trust in the provided recommendations for the three types of RSes. We have demonstrated that the context-aware RS outperformed the other two RSes in terms of accuracy, trust and other economics-based performance metrics across most of our experimental settings.

Categories and Subject Descriptors
H.3 [**Information storage and retrieval**]: H.3.3 Information Search and Retrieval–*Information filtering*

General Terms
Experimentation, Performance, Theory.

Keywords
Context, purchasing behavior, trust, accuracy, diversity.

1. INTRODUCTION
Companies use Recommender Systems (RSes) for several purposes. From the management perspective, RSes should help to increase sales of company products by providing useful recommendations. From a marketing perspective, the company should develop lasting relationships with the customers and increase customer trust over time. To achieve these goals, various types of RSes have been developed over the last 15 years. Among them, context-aware recommender systems (CARSs) [4] have received significant attention over the last few years. Most of the work on CARS has focused on demonstrating that the contextual information leads to more accurate recommendations and on developing efficient recommendation algorithms utilizing this additional contextual information. Little work has been done, however, on studying how much the contextual information affects purchasing behavior of customers and their trust in the provided recommendations.

In this paper, we study the key question: how CARS affect customer purchasing behavior and trust. We do it by conducting the controlled experiments with "live" customers in a real industrial setting (i.e., by doing the, so called, A/B testing). We compare performance of these context-aware recommendations with those produced by conventional content-based and random recommendations (random selected as a control group). This comparison is done in terms of the traditional performance metrics, such as accuracy and diversity of recommendations, as well as more economics- and business-oriented performance metrics, such as volumes of sales, quantities of purchased products, average prices of purchases, and levels of trust that the customers show in these recommendations. We also demonstrate that the CARS approach outperforms the other two alternatives by providing a good balance of accuracy and diversity of its recommendations that leads to the increased levels of sales *and* trust. This result is encouraging because it strikes a balance between the two aforementioned conflicting objectives: for the management to achieve higher levels of sales and the marketers to develop long-lasting and trusting relationships with the customers.

2. PRIOR WORK
Much research has been done on CARS, and [4] provides a broad overview of this area. It has been shown in [2] that contextual information matters in the sense that it can increase recommendation accuracy if deployed properly. Further, [25] proposed several alternative context-aware methods, compared them among themselves and demonstrated that context can increase recommendation accuracy.

The effect of recommendations on the purchasing behavior of customers and their trust has been extensively studied in RSes. For example, [28] argued that RSes help increase sales by converting browsers into buyers, increasing cross-selling opportunities, and building customer loyalty. However, the accuracy of recommendations alone is not sufficient to explain the purchasing behavior. Trust plays a key role. For instance, [26] found that the strength of recommendations has a positive impact on sales. However, recommendations influence shoppers' decisions only when they are perceived to be objective and credible. Since retailers have full control of what to include in recommendations and how to present them, it is natural for shoppers to discount credibility of online RSes because of potential manipulation by retailers. This perception is further fueled by anecdotal evidence of retailers manipulating the outcome of RSes [14, 24].

RecSys'11, October 23–27, 2011, Chicago, Illinois, USA.

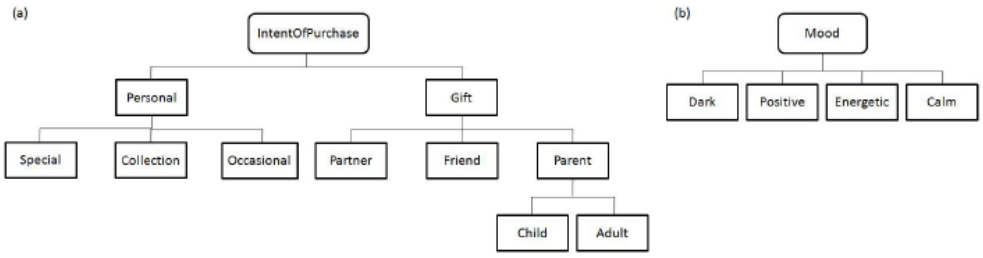

Figure 1. Hierarchical structure of contextual variables (a) intent of purchase and (b) customer's mood.

The effect of accuracy of recommendations on trust has also been studied. It has been shown that accuracy can improve trust but is not sufficient alone. For instance, [8] showed that the relevance, accuracy, completeness, and timeliness of recommendations had a significant effect on users' decision making satisfaction. [16] showed that high similarity between users and recommendations contributes to increase users' involvement, which in turn increases user satisfaction. [32, 33] showed that familiar recommendations play an important role in establishing user trust in a RS. However, [18] demonstrated that the user's familiarity with the recommendations increased trust in recommender's benevolence and integrity, but not trust in its competency. [35] showed that the way familiar and unfamiliar items are balanced in a recommendation list influences users' trust in perceived usefulness of, and satisfaction with RSes [20] stated that the accuracy of the predictions provided by a RS is only one of the possible variables affecting the service overall trustworthiness. Many scholars have demonstrated that diversity can have an important role on both trust and the economics of customers' behavior. Most researchers agree that consumers generally prefer more variety when given a choice [17, 6]. [13] demonstrated that RSes that discount item popularity in the selection of recommendable items may increase sales more than RSes that do not. Similarly, [9] showed that increased product variety made available through electronic markets can be a significantly larger source of consumer surplus gains. [23] found that diversity can provide significant gains but it has to be carefully tuned. Finally, [31] showed that higher variety seeking decreases receptivity to customized offers. Some researchers have also investigated the combined effect of accuracy and diversity. For instance, [10] demonstrated that additional recommendations of familiar products serve as a context within which unfamiliar recommendations are evaluated. [21] demonstrated that both the number of items recommended to the user and the recommendation accuracy, measured by the number of recommended items accepted by the user, had significant effects on user satisfaction. [23] demonstrated that there may be significant gains from introducing diversity into the recommendation process, but its introduction has to be carefully tuned. An important contribution is given in [1] which demonstrates the existence of a trade-off between accuracy and diversity. Ranking recommendations according to the predicted rating values provides good predictive accuracy but it tends to perform poorly with respect to recommendation diversity. Finally, [26] have investigated the effect of RSes on the price that customers want to pay for the recommended items and found that providing value-added services, such as recommendations, allows retailers to charge higher prices. All this prior work focuses on examining relationships between accuracy, diversity, trustworthiness of recommendations and increased levels of sales and other economic indicators for the *traditional* recommender systems. To our best knowledge, no research has been done on studying these effects for the *context-aware* recommender systems, especially in the context of conducting controlled live experiments in real industrial settings. In this paper we focus on these issues for the CARS.

3. METHODOLOGY

We conducted an experiment in a real-world setting in partnership with a well-known Italian firm operating in the publishing industry worldwide. The company's Web division mainly sells comic books and related products, such as DVDs, stickers, and T-shirts. As a part of its normal business, the company sends a weekly non-personalized newsletter to approximately 23,000 customers and agreed to send personalized recommendations of comic books via e-mail to a sample of this customer base as a part of our project. According to the privacy laws, the firm asked customers to state explicitly if they wanted to join this project to improve the customer service. The final number of customers who agreed to participate and provided enough responses for our study was 260, corresponding to the participation rate of about 1%. Our personalized newsletter was sent to the participating customers in addition to the traditional weekly newsletter.

In this study, we compared the performance of a content-based and a context-aware RSes. We also used an RS producing random recommendations as a control group. The 260 study participants were randomized into the three experimental treatment conditions, each of the three groups receiving either random, content-based, or context-aware recommendations. The average response rate (i.e., users who gave feedback during the experiment) was about 65% for each treatment condition. In the first week, as the initial step of our experiments, we asked the participants to rate a representative set of twelve comics selected by the firm in order to build the initial user profiles. This set of comic books was representative of the whole item database and it was the same for each user. After that, each subject received a personalized weekly newsletter displaying 10 recommended comic books for 9 consecutive weeks. It contained a link to a personal recommendation page displaying the ten recommended items. Five were "recommended brand new items" selected from brand new arrivals at the firm (about 30 brand new published comic books per week), and five were "recommended old items" selected from the arrivals in the past two months (about 250 items). Each item was presented with the following information: title, cover image, description, a "see more details" link. The customers were invited to rate each recommended product by clicking on a (1-5) point scale. These ratings were used to update the*UserProfile(i)* for each user as described in Section 3.1. At the end of the experiment we provided participants with the final survey in which we asked each of them 11 questions presented in Table 1. The purpose of these questions was to measure how much the participants trusted the received recommendations. The questions were composed according to the literature on experimental design and on trust [22, 7, 34, 29, 11].

Table 1. Table captions should be placed above the table

Measure		Question in the survey
Check	Q_1	I usually trust people
Ability	Q_2	This personalized newsletter is like a real expert in assessing comic books
	Q_3	Personalized newsletters provided me with relevant recommendations
	Q_4	Personalized newsletters recommended comic books that I didn't know
	Q_5	I am willing to let this newsletter assist me in deciding which product to buy
Integrity	Q_6	The newsletter is reliable
	Q_7	I trust the personalized newsletter
Benevo-lence	Q_8	The company created the personalized newsletter to help me
	Q_9	The personalized newsletter is a service provided by the company to customers
Offline	Q_{10}	I bought some of the recommended products offline
Price	Q_{11}	I think the recommended products were expensive

3.1 Description of the recommenders

During the experiment we used three different RSes, a content-based recommender, a context-aware recommender and a random recommender. We used the content-based recommender as a "benchmark" and we chose this recommendation algorithm (as opposed to a collaborative filtering) because the experiment was carried out with a relatively low number of participants. Given the sparsity of the user/item matrix, it would be very difficult to generate meaningful recommendations by using a collaborative engine.

3.1.1 Content-based

The content based algorithm simply recommends items that are similar to the ones the users preferred in the past [27]. As defined in the literature [3], this algorithm computes rating *u(i,s)* of item *s* for user *i* based on the ratings $u(i,s_j)$ assigned by user *i* to items $s_j \in S$ that are similar to item *s*. In particular, let *ItemProfile(s)* for item *s* and *UserProfile(i)* for user *i*, be two vectors representing the item characteristics and the customer preference, respectively. *ItemProfile(s)* are computed by extracting a set of keywords from a description of item *s*. The keywords describe the item and its contents, including author and publisher details. *UserProfile(i)* is computed by analyzing the content of the items previously seen and rated by user *i*. In particular, the vector is defined as a vector of weights $(w_{i1}, \ldots, w_{iz})$, where each w_{ij} denotes the importance of keyword *j* to user *i*. We computed w_{ij} as an "average" of the ratings provided by user *i* to those items that contained the keyword $j \in Z$. In our study, we assumed that $z = 80$, thus restricting the keyword profile lengths to 80 words. Candidate items are compared with user profile and the most similar items are recommended. We compute relevance *u(i,s)* of item *s* to user *i* by matching the *UserProfile(i)* and the *ItemProfile(s)*. The top 10 items with the highest score are presented (recommended) to the user in the newsletter. Since we adopt a content-based engine which uses item features, we checked that each item had the same amount of information (i.e., title, sub-title and description) in order to avoid introduction of any biases (e.g., recommending items with long descriptions more often than items with short descriptions, or vice versa).

3.1.2 Context-aware

Since our aim was to fairly compare a traditional (content-based) RS with a CARS, the CARS developed for our experiment used the same content-based algorithm discussed in the previous section. The only difference is that we used *UserProfile(i,k)* which is the profile of user *i* in context *k* (e.g., a gift for a parent in Fig. 1(a)) instead of *UserProfile(i)* which does not consider the context *k*. We computed profile *UserProfile(i,k)* by following the pre-filtering approach [4, 25] by analyzing the content of the items previously seen and rated by user *i* in context *k*. In particular, the contextual information *k* is used as a label for filtering out those items that were not rated in this context *k*, i.e, this method selects from the initial set of all the ratings *only* those referrings to context *k*. As a result, *UserProfile(i,k)* contains only the data pertaining to context *k*. After that, the content-based algorithm is launched on *only* this selected data to produce recommendations specific to context *k*. We follow the representational approach to defining contextual information [12]. In particular, we follow [4, 25] by defining it with a set of *contextual attributes (variables)* as follows. First, we assume that domain of contextual attribute *K* is defined by a set of *q* attributes $K = (K_1, \ldots, K_q)$ having a hierarchical structure associated with it. The values taken by attribute K_q define finer levels, while K_1 coarser levels of contextual knowledge [19]. In our experiment, we used two distinct contextual variables: the "intent of a purchase" made by a customer and the "customer's mood". These two variables are presented in Figure 1. The "intent of purchase" contextual variable distinguishes whether the user is looking for recommendations for his/her personal interest (further distinguished between recommendations for his/her collections, special issues or occasional reading) or for a gift (further distinguished between recommendations for a gift to a partner, a friend, etc.). Contextual variable "mood" distinguishes between different moods of the customer who may be looking for different recommendations depending on his/her type of the mood which can be dark, energetic, positive or calm in our study. We selected "intent of purchase" and "mood" contextual variables in our study after setting up focus groups and discussing the results produced by focus groups with the management. We also used other recommendation applications, such as music recommendations, as reference points for identifying contextual variables. When users of the contextual treatment group received the newsletter, it was asked them to specify the context in which they wanted to receive recommendations, (i.e., for a personal purpose or for a gift and then for whom or what was their mood before showing them the recommendations list). Then recommendations only for the specified context were shown to the participants. It was possible to change the target context once it was set, therefore customers could see and rate the recommended items also in another context.

3.1.3 Random/Control group

Unlike the content-based and context-aware approaches, the random approach does not take the user profile into consideration when recommending new products. Instead it randomly selects, without replacement, a set of items to recommend from the products that have not been recommended or purchased in before.

3.2 Performance metrics

We measured the accuracy and diversity of the recommendations received by the customers in the three groups, as well as their purchasing behavior and trust. Accuracy was measured by precision and average ratings. Among the traditional IR performance metrics, such as precision, recall and F-measure, only precision could be computed in our case, since it was not possible to know the ratings of the unseen items needed to compute the recall and the F-measure. According to [15], precision was measured as:

$$P = \frac{N_{rs}}{N_s} \quad (2)$$

where N_s is the total number of the items recommended to the customer ("selected" by the RS as items to be recommended) and N_{rs} is the number of items which proved to be "relevant" ("good recommendations") for the customer among those selected by the RS. We considered an item being "relevant" if it was rated as 3, 4 or 5 on the 0 – 5 scale. We decided to consider items rated as 3, 4 or 5 as relevant instead of considering only items rated as 4 or 5 as discussed in [15] since our rating scale was from 0 to 5 instead of 1-5 scale (as in [15]). We also measured accuracy as the percentage of positive ratings over time, i.e., as the percentage of users providing an average rating greater than 3 in each week. The percentage of users with a positive average rating in week z was computed as

$$Users\ with\ positive\ ratings_z = \frac{users\ with\ Average\ rating_{z,u} > 3}{total\ number\ of\ users} \quad (3)$$

where $Average rating_{z,u}$ is the average rating provided by user u in period z over the items rated, $rating_{n,z,u}$ is the rating provided by user u to item n in period z and n is the total number of items rated by u in period z:

$$Average\,\text{rating}_{z,u} = \frac{\sum_n rating_{n,z,u}}{n} \quad (4)$$

We measured diversity of recommendations using entropy [30]. We used four comic book categories, according to the main classification the company uses to present its products in the web-site: 1) Marvel comics (including the well-known comic books popularized by the American publisher); 2) Manga comics (including all comic books published in Japan); 3) other comics (including all comic books popularized by either European publishers or American publishers other than "Marvel" brand); 4) bundled comics (including any kind of comic books sold in association with a DVD or other media contents). Entropy was computed as

$$Entropy(X) = \sum_i P(X_i)\log_2(P(X_i)) \quad (5)$$

where *Entropy(X)* is the uncertainty (or inconsistency) of variable X (or the categories), and $P(X_i)$ is the probability that comic book X belongs to category i. We used three measures to represent purchasing behavior of users before and during the experiment. We measured the purchased "quantity" per month per capita in each group by counting the number of products bought in each group divided by the number of months divided by the number of customers in each group. We also measured the average "price" of the products bought by computing the average of the prices of the products bought in each period by each treatment group. As previously mentioned, in addition to evaluating user feedback in terms of ratings and purchases, we gathered additional feedback from participants to study whether there were differences in customers' trust across the treatments. The questions are reported in Table 1. Each answer was provided on the (1-5) scale. The first question (Q_1) was used to check possible biases in the responses. The questions from Q_2 to Q_5 are measures of "ability", the next two (Q_6 and Q_7) are measures of "integrity", the next two (Q_8 and Q_9) are measures of "benevolence". The last two questions, Q_{10} and Q_{11}, were used as measures of offline purchases and perceived price, respectively. The constructs for trust were derived from prior studies. We limited the scope of the survey to testing trusting beliefs which consist of three constructs: ability, benevolence, and integrity [22, 7]. We adapted a previously used set of questions and scales from [11, 29, 34] where trusting beliefs were also linked to recommendations. We selected and adapted the items that were relevant to our application context.

4. RESULTS

In this section we describe the results of the comparison of customer responses to our recommendations across various experimental setting reported in Section 3. The average response rate across all the participants over the entire period was 58.9%.

Table 2. Purchasing behavior of the three groups

		Content-based	Context-aware	Control
Sales	*before*	2.03	1.95	0.91
(€)	*during*	2.38	2.50	0.94
%var		+16.9***	+28.2***	+3.6**
Quantity	*before*	0.33	0.37	0.15
(# items)	*during*	0.45	0.34	0.10
%var		+36.8**	-7.7***	-31.9**
Price	*before*	6.18	5.26	6.20
(€/item)	*during*	5.29	7.31	9.43
%var		-14.5***	+38.9***	+52.1***

***Significant at p<0.01. **Significant at p<0.05.

Table 2 reports the purchasing behavior of the customers in the three treatment groups (content-based, context-aware and control) before and during the experiments. In order to make a meaningful comparison, the firm gave us access to the data pertaining to the purchasing behavior of the customers involved in the experiment in a period of twenty months before the experiment beginning. As Table 2 shows, the two groups that received personalized recommendations modified their behavior by increasing the money spent per month per customer. The increase in the context-aware group is higher than that of the content-based group, being 28.2% and 16.9% respectively. The money spent remained almost the same in the control group (3.6%). If sales are decomposed into quantity and price, we observe that the behavior change in the content-based group is caused by an increase in the quantity bought (36.8%) and a decrease in the price of the items (-14.5%). On the contrary, the quantity decreases for the context-aware group by 7.7% while the average price of items increased by 38.9%. The quantity decreases by 31.9% in the control group while the few items bought have higher price (52.1%). The statistical differences were tested using the Wilcoxon test [5], and the statistically significant differences are marked with asterisks (*) in Table 2. These results will be explained and further discussed in the next sections. Table 3 reports the average answers to the survey per each treatment group. Numbers smaller than 3 mean that customers mistrusted the system, and vice versa for numbers greater than 3. Again, statistically significant differences

Table 3. Results of the final survey: measures of trust and additional metrics

	check	ability				integrity		benevolence		offline	price
	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	Q_{11}
Content-based	2,889	3,231	3,404	2,731**	2,745	3,231	3,020**	3,500	4,118	3,712**	2,904
Context-aware	3,125	3,169	3,458	3,056	3,141	3,417	3,222	3,694	4,028	3,620	2,957
Control	3,054	3,083	3,351	3,361**	3,200	3,611	3,472**	3,429	3,943	3,114**	2,857

** Differences between values in the same column are statistically significant with p<0.05.

were marked with asterisks (*) in Table 3. The answers to Q_1 were not statistically significantly different across the groups, so there was no bias in the group composition. Looking at the measures of trust (Q_2 to Q_9), the customers in the content-based group slightly mistrusted the system ability to let them discover new items (Q_4) and the system ability to assist the customer (Q_5). Their trust in the integrity of the newsletter (Q_7) was neutral. In the other two groups there is no measure showing mistrust by customers. Customers in the context-aware group showed a neutral perception of trust for a measure of ability (Q_4, discover of new items). The customers who received random recommendations showed a neutral perception of trust for a measure of ability (Q_2, the newsletter is an expert). The remaining two questions (Q_{10} - offline purchases and Q_{11}, - price) show that all customers bought some products offline and that they did not perceive that the purchased items were expensive. We found statistically significant differences only between the control and the content-based groups for a measure of ability (Q_4), a measure of integrity (Q_7), and the answer for offline purchases (Q_{10}), with $p<0.05$. This means that the customers in the control group believed that the system let them discover new items significantly better than those in the content-based case. The trust in the integrity of the newsletter by the customers in the content-based was significantly lower than that of customers in the control group. Finally, customers receiving content-based recommendations stated they purchased significantly more comic books offline compared to those in the control group.

After presenting the effects of recommendations on customers' purchases and trust, we now show how accurate and diverse the recommendations generated by the different systems are. As explained in Section 3, we analyzed accuracy and diversity of recommendations received by customers while studying customers' behavior. As also mentioned in Section 3, we used two measures of accuracy, namely precision and percentage of positive ratings. Figure 2 reports the precision of recommendations for each group, as defined in Section 3.2, during the ten weeks of the experiments.

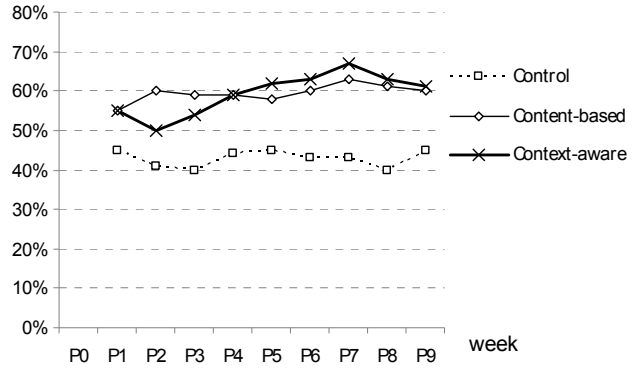

Figure 2. Precision of recommendations in the three groups

As Figure 2 demonstrates, the precision of the recommendations generated by the two personalized RSes is significantly higher than that of the random recommendations. The precision of the content-based RS and that of the CARS across the ten weeks is similar. Further, we found no statistically significant difference between the two groups. However, the precision of the CARS is slightly greater than that of the content-based RS after the fourth week. The reason is that the CARS approach takes more time to learn the preferences of customers because of the sparsity of the user-item matrix which is computed for each context with smaller amount of data; and then it outperforms the content-based approach after the initial learning is complete. The results observed for the precision in Figure 2 are reinforced by the similar results reported in Figure 3 for a different measure of percentage of positive ratings (i.e., percentage of customers who provided positive ratings (greater than 3) on average in each week). As in Figure 2, in this case the CARS method performed slightly better than the content-based one starting from week 4, as a higher percentage of customers provided positive ratings to the recommended items. The performance decrease observable in the last three weeks was caused by the technical need of keeping the computational time low. To that aim, we cut the quantity of items that can be recommended to customers thus reducing the amount of time necessary to compute the list of recommended items.

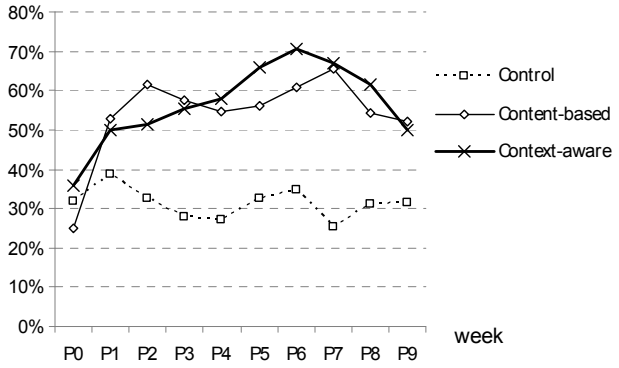

Figure 3. Percentage of customers providing positive ratings

The diversity of recommendations can be measured by the entropy defined in (5) and by the response to Q_4. Figure 4 reports the entropy of the recommendations received by the customers in the three groups. It is measured for each customer by considering the whole set of recommendations received during the ten weeks of experiment. The graph reports the percentage of customers who received a set of recommendation with a certain level of entropy. It is interesting to notice that the diversity of the CARS was very similar to that provided by a random recommender, while the recommendations generated by the content-based RS were much less diverse. An explanation of the fact that the CARS is almost as diverse as the random recommender is that the user can change the target context thus reaching different parts of the space of products.

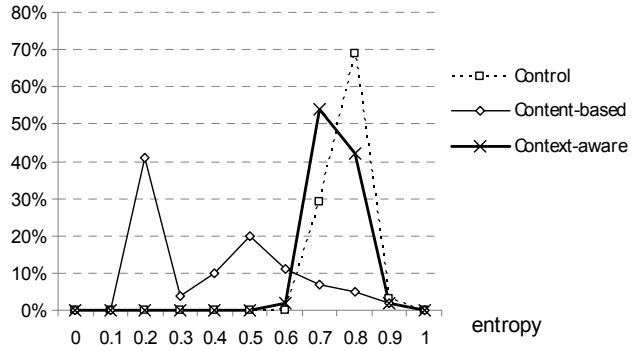

Figure 4. Diversity of recommendations measured by entropy

An alternative way to measure the diversity of recommendations is via the answers to Q_4 in Table 3. In fact, Q_4 asked the customers whether they agree with the statement that "personalized newsletters recommended comic books that I didn't know". This means that Q_4 can be considered as a proxy for the diversity measure. The results in Table 3 are consistent with those reported in Figure 4: random recommendations are perceived as the most

diverse while the content-based ones are the least diverse. In order to better visualize the results and make the discussion easier, Figure 5 plots the three different RSes we used in the experiment according to the average accuracy and diversity of the recommendations they generated. For each system we computed the average precision over the ten weeks (x-axis) and the average entropy over the distribution in Fig. 4 (y-axis). The content-based recommendations were characterized by high accuracy and low diversity, whereas the random recommendations were highly diverse but inaccurate. The context-aware recommendations were as accurate as those generated by the content-based RS and their diversity was almost equal to those generated randomly. This means that the CARS method dominates the other two alternatives when both accuracy and diversity measures are considered. In order to find additional evidence for these observations, we built several statistical models between the variables used in the experiments to measure accuracy and diversity and the variables used to measure purchasing behavior and trust. For the sake of conciseness, we only display the most relevant models among all those that were built. In Table 4 all the variables used in the analysis are shown (grouped by the concept they measure).

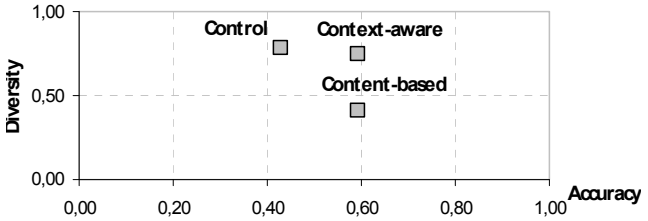

Figure 5. Accuracy and diversity of the recommender systems

The answer to Q_4 ("The personalized newsletter recommended comic books that I did not know") was used only as a measure of diversity. As Table 2 shows, we also measured the purchasing behavior of the customers in the three treatment groups before the experiment through the quantity and the price of the product bought. However, we did not use these variables in the statistical models.

Table 4. Variables used in the analysis of results

Accuracy	*Rtng*	Average rating provided by the customer
	Prec	Average precision of recommendations
Diversity	*Entr*	Entropy
	Q_4	Answer to Q_4
Purchasing behavior	*Qty*	Purchased quantity *during* the experiment
	Price	Average price *during* the experiment
Trust	Q_2-Q_9	Answers to Q_2 to Q_9 (except Q_4)

As a first step, we investigated the relationship between *Qty* and the measures of accuracy, diversity and trust. We found that *Qty* can be explained by accuracy alone. A linear regression model built with *Qty* as dependent variable and *Prec* as independent variable was significant at $p<0.05$. Moreover, *Qty* can be explained by trust alone. In particular, two models were significant at $p<0.05$, one using Q_3 (trust in the "relevance" of recommendations) as measure of trust, one using Q_6 (trust in the "reliability" of recommendations). On the contrary, *Qty* cannot be explained by any measure of diversity alone. By combining the above cited measures we obtained the model shown in Table 5, which reports the coefficients of a linear regression (the standard errors are in parentheses). The model shows that the quantity purchased by customers can be explained by a combined effect of accuracy and diversity, with a stronger effect of accuracy. Although these statistical relationships do not necessarily mean a "causal" relationship, the model indicates that delivering more accurate recommendations can make people buy more products, while diverse but inaccurate recommendations do not have this effect.

Table 5. Linear regressions among quantity purchased, precision and trust in "reliability"

Independent variables:	Dependent variable: *Qty*
Prec	4.195 (2.096)**
Q_6	.822 (.429)**
Constant	-3.954 (1.740)**

** Significant at $p<0.05$. The number of observation is 162.

We then built models to explain trust by accuracy and diversity. We used ordinal probit models because of the characteristics of the variables that measure trust (Q_2 to Q_9). We decided to measure trust by Q_6 (trust in the "reliability") because this variable proved to be the best to combine all the results. Similar results were obtained by using Q_7 (trust in the newsletter). We found that trust can be explained by diversity alone, namely by using Q_4 as independent variable, with very high significance ($p<0.001$). Trust can also be explained by accuracy alone, namely by using *Rtng* as independent variable, with a lower significance ($p<0.05$). We then built an ordinal probit model which combines accuracy and diversity. Table 6 shows the coefficients of the model (standard errors are in parentheses). This models suggests that accuracy is not enough to explain trust. Trust can be better explained by a combination of accuracy and diversity, with a stronger and more significant effect of diversity. The ordinal probit model produced four constant values (the fifth is redundant as Q_6 takes five values), all significant at $p<0.01$ except one.

Table 6. Ordinal probit model among trust in "reliability", precision and trust in "discover of new item"

Independent variables:	Dependent variable: Q_6
Rtng	.257 (.110)**
Q_4	.414 (.066)***

***Sig. at $p<0.001$. **Sig. at $p<0.05$. Number of observation 162.

Again, although we cannot state that a "causal" relationship exist among these variables, the model suggests that customers tend to trust the recommendations, particularly their reliability, when they are diverse. Although accuracy has an effect on trust, it cannot increase trust in the absence of diversity. If recommendations are accurate but not diverse, customer may distrust them. On the contrary, delivering diverse but inaccurate recommendations can be sufficient to increase trust. Finally we built models to explain the willingness of customers to spend more money for individual products. We used *Price* as dependent variable in linear regression models, and the measures of accuracy, diversity and trust as independent variables. We did not find any significant statistical relationship between *Price* and any variable representing accuracy or diversity. The only variable which proved to be able to explain *Price* is trust, namely Q_6. The model is described in Table 7. This model suggests that people are willing to spend more money for an individual item if they trust the system rather than because of the accuracy of the recommendations. Because of the relationships previously commented, we can deduce that only if the combination of accuracy and diversity is such that trust increases, customers will buy more expensive products. By combining the findings described so far, we can draw Figure 6 and Figure 7, where the different effects of accuracy and diversity on trust and on the purchasing behavior are shown. The model explains the behavior that we observed in the experiment by using different RSes.

Table 7. Linear regression between average price of purchased items and trust in "reliability"

Independent variable:	Dependent variable: *Price*
Q_6	.245 (.102)**
Constant	-.524 (.358)

** Significant at p<0.05. The number of observation is 208.

If accuracy is high and diversity is low, then trust is moderate, the purchased quantity is high and the average price of purchased products low. In fact, the customers receiving content-based recommendations distrusted certain aspects of recommendations but increased the purchased quantity at a low price. If accuracy is low and diversity is high, then trust is high and the price of products bought is also high, but the purchased quantity is low. In fact, the customers who received random recommendations trusted them, bought very few items but at a higher price than they used to do before the experiment. Finally, if both accuracy and diversity are high, then trust is high and both quantity and price are expected to be high. The customers who received context-aware recommendations, which were both accurate and diverse, trusted the system more than in the case of the content-based RS (although not as much as for the random RS), and the products they bought were more expensive (higher priced). The model in Figure 6 would predict an increase in both quantity and price. Actually, in the context-aware RSes the sales increased while quantity slightly decreased. This behavior can be explained by the intuitive observation that people do not want to increase their expenses drastically. In our experiment, these customers decided to limit quantity but, nevertheless, the money they spent increased significantly compared to their past behavior.

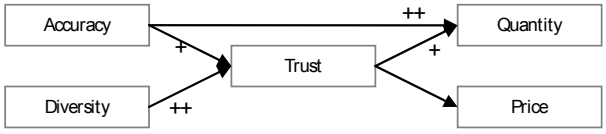

Figure 6. Combined effects of accuracy and diversity on purchasing behavior and trust

These models also explain the results shown in Table 2 where sales are disaggregated in quantity and price for the three systems. Figure 7 summarizes the behavior explained by the statistical models. The context-aware system, which dominates both the content-based and the random one in terms of accuracy and diversity, showed the best properties: even though the quantity of the products purchased slightly decreased, the average price increased and trust remained high with respect to the other systems. As a consequence, the system provided the best results in terms of increase in sales and trust, as was explained above.

5. DISCUSSION AND CONCLUSIONS

This research aims at measuring of how purchase behavior and trust of customers is affected by different types of recommendations, including those generated by CARS, and by such factors as accuracy and diversity of recommendations. The measurements performed in this study were done using "live" controlled experiments (A/B tests) in partnership with a well-known Italian publishing firm. One of our findings, supporting prior observations [26, 18, 20], is that accuracy of recommendations alone does not explain economics of purchasing behavior of customers and also does not explain how much they trust these recommendations. Therefore, we also considered diversity of recommendations, besides their accuracy, and compared performance of some of the context-aware, content-based and random recommendations in terms of these two measures. Our results are summarized in Figures 5 and 6. Figure 5 demonstrates that CARS outperformed the content-based system considered in the paper, as well as the control case of random recommendations along the dimensions of accuracy *and* diversity.

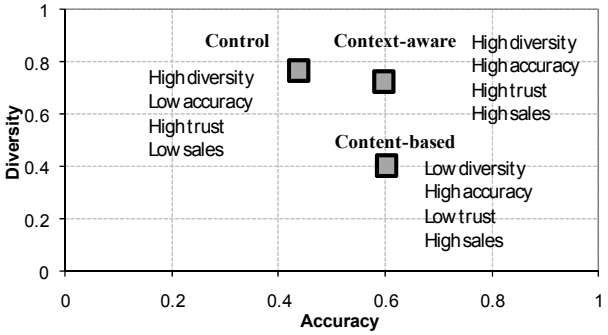

Figure 7. Effects of accuracy and diversity

Figure 6 shows the combined effects of accuracy and diversity of recommendations on trust and on the customer purchasing behavior, as measured in terms of the quantity of purchased products and the average price they are willing to pay for these products. In particular, it shows that trust is being affected by both accuracy and diversity of recommendations, diversity being the dominant force here. Also, we show that the quantity of the purchased products is affected by the accuracy of recommendations and by trust, accuracy being the dominant force in this process. Finally, we show that the average price of a purchased product is being directly affected by the trust the customer has in the recommendation, i.e., the more the customers trust the recommendations, the more they are willing to pay for the products. If the results of Figures 5 and 6 are combined, then we conclude that the CARS systems produce better recommendation outcomes than the content-based and random systems considered in the paper in terms of customer purchasing behavior *and* trust in the provided recommendations. These results have practical importance for the industry because management and marketing divisions of companies have somewhat conflicting goals: management wants to increase sales of its products and the resulting profits, while marketing is focused on building lasting relationships with the customers and increasing customer trust over time. Our study demonstrates that, unlike some alternative approaches (such as content-based and random), a CARS can provide a good balance of accuracy and diversity of recommendations that result in the increased levels of sales *and* trust. And we demonstrated these effects in our "live" controlled experiments on a real-life application. All this provides additional important evidence for the usefulness and practicality of CARS and the necessity of their deployment in various applications.

As a part of the future work, we would like to test the results reported in this paper on other types of recommendation applications and for other types of industries. This should allow us to generalize and broaden our conclusions and perhaps identify additional factors affecting economic behavior and trust of customers besides the accuracy and diversity of recommendations studied in this paper. We would also like to conduct a bigger study involving more customers than we currently used. Finally, we would like to compare our results with different diversification techniques across a broader range of RSes than we used in this study in order to deeply test the trade-off between customer trust and recommendations diversity.

6. REFERENCES

[1] Adomavicius, G. and Kwon, Y. O. 2010. Improving aggregate recommendation diversity using ranking-based techniques. Forthcoming on *IEEE Trans. Kn. Data Eng*.

[2] Adomavicius, G., Sankaranarayanan, R., Sen, S. and Tuzhilin, A. 2005. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Trans. Inf. Sys.* 23, 103-145.

[3] Adomavicius, G. and Tuzhilin, A. 2005. Towards the next generation of recommender systems: A survey of the state-of-the art and possible extensions. *IEEE Trans. Kn. Data Eng*. 17, 6, 734-749.

[4] Adomavicius, G. and Tuzhilin, A. 2011. Context-Aware Recommender Systems. In *Handbook on Recommender Systems*, Ed. Springer.

[5] Barnes, J.W. 1994. Statistical Analysis for Engineers and Scientists. McGraw Hill, Singapore, 1994.

[6] Baumol, W. E. and Ide. A. 1956. Variety in retailing. *Man. Sci.* 3, 1, 93-101.

[7] Beldad, A., de Jong, M., Steehouder, M. 2010. How shall I trust the faceless and the intangible? A literature review on the antecedents of online trust. *Computers in Human Behavior* 26, 5, 857-869.

[8] Bharati, P. and Chaudhury, A. 2004. An Empirical Investigation of Decision-Making Satisfaction in Web-Based Decision Support Systems. *Dec. Sup. Sys.* 37, 2, 187-197.

[9] Brynjolfsson, E., Smith, M. D. and Hu, Y. 2003. Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers. *Man. Sci.* 49, 11, 1580-1596.

[10] Cooke, A. D. J., Sujan, H., Sujan, M., Weitz, B. A. 2002. Marketing the unfamiliar: The role of context and item-specific information in electronic agent recommendations. *J. of Marketing Res.* 39, 4, 488-497.

[11] Doney, P. M., and Cannon, J. P. 1997. An examination of the nature of trust in buyer–seller relationships. *J. Marketing*, 61, 35-51.

[12] Dourish, P. 2004. What we talk about when we talk about context. *Personal and Ubiquitous Computing* 8, 19-30.

[13] Fleder, D., and Hosanagar, K. 2009. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Man. Sci.* 55, 5, 697-712.

[14] Flynn, L. J. 2006. Like This? You'll Hate That. New York Times, Jan. 23, 1.

[15] Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. 2004. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Sys.* 22, 5-53

[16] Hess, T. J., Fuller, M. A., and Mathew, J. 2005. Involvement and Decision-Making Performance with a Decision Aid: The Influence of Social Multimedia, Gender, and Playfulness. *J. Man. Inf. Sys.* 22, 3, 15-54.

[17] Kahn, B., and Lehmann D. R. 1991. Modeling choice among assortments. *J. Retailing* 67, 3, 274-299.

[18] Komiak, S., and Benbasat, I. 2006. The Effects of Personalization and Familiarity on Trust and Adoption of Recommendation Agents. *MIS Quarterly* 30, 4, 941-960.

[19] Kwon, O, and Kim, J. 2009. Concept lattices for visualizing and generating user profiles for context-aware service recommendations. *Exp. Sys. with Appl.* 36, 1893-1902.

[20] Lenzini, G., Houten, Y.V., Huijsen, W., and Melenhorst, M. 2009. Shall I Trust a Recommendation? Towards an Evaluation of the Trustworthiness of Recommender Sites. In Proceedings of ADBIS Workshop. 121-128.

[21] Liang T. P., Lai H. J. and Ku Y. C. 2006. Personalized Content Recommendation and User Satisfaction: Theoretical Synthesis and Empirical Findings, J. *Man. Inf. Sys.* 23, 3, 45-70.

[22] Mayer, R. C., Davis, J. H., and Schoorman, F. D. 1995. An integrative model of organization trust. *Academy Man. Rev.* 20, 3, 709-734.

[23] Mcginty, L., and Smyth, B. 2003. On the role of diversity in conversational recommender systems. In Proceedings of the Fifth International Conference on Case-Based Reasoning. 276-290.

[24] Mui, Y.Q. 2006. Wal-Mart Blames Web Site Incident on Employee's Error. Washington Post, Jan. 7.

[25] Panniello, U., Tuzhilin, A., Gorgoglione, M., Palmisano C., and Pedone, A. 2009. Experimental comparison of pre- vs. post-filtering approaches in context-aware recommender systems. In Proceedings of RecSys '09, 265-268.

[26] Pathak, B., Garfinkel, R., Gopal, R. D., Venkatesan, R., and Yin, F. 2010. Empirical Analysis of the Impact of Recommender Systems on Sales. *J. Man. Inf. Sys.* 27, 2, 159-188.

[27] Pazzani, M. J. and Billsus, D. 2007. Content-based recommendation systems. In The adaptive web, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Lecture Notes In Computer Science, Vol. 4321. Springer-Verlag, 325-341.

[28] Schafer, J. B., Konstan, J. A., and Riedl, J. 2001. E-commerce recommendation applications. *Data Min. Kn. Disc.* 5, 1, 115-153.

[29] Schoorman, F. D., Mayer, R. C., and Davis, J. H. 2007. An integrative model of organizational trust: Past, present, and future. *Academy Man. Rev.* 32, 2, 344-354.

[30] Shannon, C. 1948. A Mathematical Theory of Communication. Bell system Technical Journal, 27.

[31] Simonson, I. 2005. Determinants of customers' responses to customized offers: conceptual framework and research propositions. *J. Marketing* 69, 1, 32-45.

[32] Sinha, R., and Swearingen, K. 2001. Comparing Recommendations Made by Online Systems and Friends. In Proceedings of the 2nd DELOS Workshop on Personalisation and Recommender Systems, Dublin, Ireland, June 18-20.

[33] Swearingen, K., and Sinha, R. 2001. Beyond Algorithms: An HCI Perspective on Recommender Systems. In Proceedings of the ACM SIGIR Workshop on Recommender Systems, New Orleans, LA, September 13.

[34] Wang, W., and Benbasat, I. 2005. Trust In and Adoption of Online Recommendation Agents. *J. of the AIS.* 6, 3, 72-100.

[35] Xiao B. and Benbasat I. 2007. E-commerce Product Recommendation Agents: Use, Characteristics, and Impact. *MIS Quarterly.* 31, 1, 137-209.