# Which Products Would You Like to See?

## User preferences for top-N e-commerce products recommendation in different shopping moments

Alexander Robert Kutzke, Jaime Wojciechowski, Rafaela Mantovani Fontana, Joao Eugênio Marynowski, Razer Anthom Nizer Rojas Montaño, Rafael Romualdo Wandresen and Arno Paulo Schmitz
Professional and Technological Education Department
Federal University of Parana (UFPR)
Curitiba Brazil
alexander, jaimewo, rafaela.fontana, jeugenio, razer, rafael.wandresen, arno@ufpr.br

*Abstract*— **Recommendation systems are essential tools in e-commerce web sites to help users find the items that might interest them. Although much research has shown the best algorithmic solutions, few studies rely on real products databases and users' perceptions. This study objective is, thus, to identify what are Internet shoppers preferences for Top-N products recommendation regarding the shopping steps. We have built different algorithmic solutions and presented users in the context of two different moments in the shopping process. Through a survey, 202 users evaluated the presented items and the collected data was compared using statistical and qualitative data analysis. Our results have shown that users prefer different types of recommendations in different shopping moments. This evidence generates opportunity to further research and useful data for web retailers.**

***Products recommendation; e-commerce; user perception; content-based; collaborative-filtering***

## I. INTRODUCTION

In a context of information overload, recommendations systems are tools that help Internet users retrieving the right information. These systems are responsible for predicting items that would be interesting to users [1]. Recommendations have been given to users of a variety of websites, as movies, news or academic papers. Specially in e-commerce context, these systems are an essential part of the business, as they may help improving sales and revenue [2].

An important task in recommending products to e-commerce shoppers is to find out a small set of items which might be the most appealing to users – it is named as a top-N recommendation task [3]. Although a number of studies have been verifying the accuracy and performance of algorithms to propose items to customer [3], user-centric evaluations are also necessary, as they complement findings from algorithmic tests in open databases [3; 4].

A typical sale in an e-commerce comprises a number of shopping steps, such as product selection, shopping cart, payment etc. The probability of conversion – that is, effectively selling the product – raises as the user progresses in these steps [5]. Considering the importance of recommendations in e-commerce environments and that shoppers objective might be different throughout this process, we aim at verifying not just user perception on received recommendations, but also whether they are different in two shopping steps: the moment when a product is selected and viewed; and the moment when a product is inserted in shopping cart.

The research question that guided this study was, thus: *"What are internet shoppers preferences for Top-N products recommendation regarding the shopping steps?"*. We aimed at identifying users' preferences for the product selection moment and for the shopping cart moment. We also evaluated content based and collaborative filtering algorithms results.

The evaluation of the composition set shown to the user still remains a topic to be investigated in research [4]. Thus, our results interest to academy, as they complement the studies focused on the evaluation of algorithms metrics – emphasizing user perception. Our results also interest to industry, considering that users recommendations have already shown to have positive impact on e-commerce sales [3; 6]. This study shows thus, preferences Internet customers may have when using online stores.

This paper is organized as follows. Next section presents related work. Section 3 shows how we conducted this study – our research approach. Our results are shown on Section 4, and discussions and conclusion in Sections 5 and 6, respectively.

## II. RELATED WORK

Recommender systems are designed to help users find products or items that better fit their needs in web context. They collect information about users to identify their interests, apply a learning algorithm to filter users features, and predict/recommend what kind of items users might prefer [1].

There are different approaches to process information to generate recommendations. The content-based (CB) approach recommends products or items similar to the ones user prefers. It evaluates the content of the items to generate new recommendations. The collaborative-filtering (CF) approach

recommends items to the consumer that people with similar preferences have liked before, generating variety in the recommendation. There is also the possibility of combining both approaches in hybrid ones [1; 7].

Academics of recommender systems have been long proposing algorithms and evaluating them. One of the approaches is to identify recommendations that fulfill users expectations at a first glance. This is the top-N recommendation, in which the algorithm must propose a few items that might interest users [3].

The study by [8] shows, for example, the importance of diversity in top-N recommendations. They have used training datasets to propose an algorithm that includes diversity in recommendations that are usually too similar. They also consider the importance of a user-centric evaluation and propose a methodology that considers the utility of the recommendation, and not just accuracy as most studies do [8, p. 2].

Another proposal that considers diversity in top-N recommendations is the one by [9]. Their solution is to generate content-based and collaborative-filtering candidates separately for a specific group of users. The items are, then, aggregated according to users' preferences and merged. The result is that content-based and collaborative filtering results are mixed together. They evaluated their results with experiments with real users, similarly to our proposal.

The researchers in [10] went beyond diversity and introduced some dynamism in their recommendations, which they call "cycling" and "serpentining". Cycling demotes some items when they are shown several times and serpentining spreads best recommendations across several pages. They identified that combining two approaches was not interesting and, sometimes, individual approaches generated a perception on users that was not clear.

The work by [11] studied how different collaborative filtering algorithms triggered satisfaction perceptions on users. They identified that the diversity generated by different algorithms did not create different perceptions, and that users seem to prefer to see similar products on recommendations.

Besides finding out those few items that interest to users, recent research has identified that the order these products are shown also influence users decisions [12]. The authors in [12] applied a user satisfaction measure to verify that order has a positive effect on recommendations.

Top-N recommendation is a relevant topic on ecommerce environments. Recent research has shown that systems recommendations have impact on sales [6] and identifying a few items that interest users might mean an increase in revenue. Our work complements the current studies by showing that not just the similarity or diversity of presented items might trigger interest in users. Shoppers expect to see different types of recommendations depending on the shopping step where they are, as shown in next sections.

## III. Research Approach

The objective of this study was to identify whether users point out different preferences for products recommendation, regarding the shopping moment where they are. We conducted the research focusing on the question: *"What are internet shoppers preferences for Top-N product recommendation regarding the shopping steps?"*. To answer that, we conducted a survey followed by a statistical data analysis and an additional automatic qualitative analysis.

According to [13], a survey collects information about individuals to contribute to the general body of knowledge in a specific field. Our focus in this study was the exploratory survey, which allows gaining insight on a specific topic and providing the basis for more in-depth research.

In the first step to conduct the survey, theoretical domain should be translated to the empirical domain, so that questions reflect the theory to be measured [13]. In this study, we have conceived the questions in a way that respondents should evaluate lists of product recommendations. The products we have shown were furniture from a real web retailer from Brazil. Based on a specific product we have predefined, we presented to users 4 products recommendation lists, in the following way:

1. We created a scenario in which the respondent should imagine she would buy a specific furniture. We presented the furniture name, description and image.
2. We asked the respondent to analyze four lists of products. We built the recommendations based on the hypothetical product selection in step 1. Each list showed six products generated as the result of the application of an algorithm or a combination of two algorithms:
    - One of them, showed only similar products, based on a clustering algorithm. As it is based on a content-based algorithm, we call it a CB-generated list;
    - The other showed only related products, based on an association rules algorithm. As it is based on a collaborative filtering algorithm, we call it a CF-generated list;
    - The other was a combination of both algorithms, showing first three products generated by the clustering algorithm and, next, three products generated by the association rules algorithm. It was a hybrid recommendation, which we call here CB+CF; and
    - The last list also combined two algorithms, but showed first the products generated by the association rules algorithm and, next, the products generated by the clustering algorithm. It is another hybrid list, which we call here CF+CB.

For each list, we asked the respondent to give us a rating for the recommendation received, from 1 to 5. The rating 1 meant "completely dislike", 2 meant "dislike", 3 meant a "neutral opinion", 4 meant "like" and 5 meant "completely

like" this products recommendation list. All respondents saw the same products in each list, but the sequence in which the lists appeared to each user was different, to reduce bias in the evaluation.

Respondents evaluated the lists in two simulated scenarios: considering that the user was in the product selection moment, and considering that the user was in the shopping cart moment.

To reduce bias on product preferences, these lists were shown for three different products: a wardrobe, an office-desk and a kitchen cabinet. All products had a description and an image provided by the web furniture retailer.

When users finished ranking products for a given moment, we presented users open-ended questions. At the product selection moment, we asked them which products recommendations they would like to see when searching and selecting a product. At the shopping cart moment, respondents were asked to comment on the products recommendations they would like to see in the shopping cart moment.

All these scenarios were presented to respondents in a website we created. The language was Brazilian Portuguese. This website also collected the responses and saved in a database. We piloted-tested the questionnaire with five regular Internet shoppers, who gave suggestions to improve products presentation and text. For the data collection, the sampling strategy was the snowball technique [14]: we spread the research call to our contacts using social media, asked them to spread to their own contacts and so on.

### A. *The Content-Based and Collaborative Filtering Algorithms*

The product lists shown to our respondents were fixed. They were previously created with the application of two different algorithms. A CB one, the clustering, and a CF one, the association rules algorithm.

The clustering solution was implemented using the K-Means algorithm [15] from the Weka environment [16]. This algorithm works by analyzing the items properties and grouping them according to the similarity of these properties. In our implementation, we used the products database provided by the web retailer with the information: product category, product sub-category, color and style. A file with 72,921 real products was used as input to the K-means algorithm to identify clusters of similar products.

Before using the resulting clusters, we simulated different scenarios to identify the best amount of clusters. We performed tests with 1,000, 1,500 and 3,000 clusters. As we aimed at having six recommended products, the amount of clusters that returned this situation was 1,500. The simulations with values lower than this resulted with too many products for each cluster, and simulations with a value greater than this resulted in less than six products per cluster.

The association rules algorithm was built using the same web retailer database, but now considering sales, and not only products properties. We used the APRIORI association rules algorithm [17; 18). It identifies products that were bought together. Our implementation was based on [19] to generate a frequent itemset. According to the APRIORI algorithm, an itemset is considered frequent when the number of transactions – in our case, sales – that contain that itemset is greater than a threshold.

The sales database had 139,460 records and we created 83,516 sales records. In each sale record there was a group of products bought together in a specific sale. These sales records were used as input to the APRIORI algorithm, with a threshold of 0.1%, which generated 389 frequent 1-itemsets, 107 frequent 2-itemsets, 68 frequent 3-itemsets, 28 frequent 4-itemsets, 8 frequent 5-itemsets and 1 frequent 8-itemset.

When a specific product was given, the recommendation created by the APRIORI algorithm returned six products that appeared together with the input product in the generated itemsets.

### B. *Data Analysis*

As we wanted to identify whether there was difference in the preference of products shown in the moment of product selection and in the moment of shopping cart, our data analysis was based on two hypotheses and three supporting questions to verify them:

- H1: On selecting a product, people do prefer to see first similar products.
    - What is the algorithm that generated the best list on product selection?
    - What is the algorithm that generated the worst list on product selection?
    - Is there a difference on users' perception when products list contains mixed algorithms results on product selection?
- H2: On shopping cart, people do prefer to see first different products.
    - What is the algorithm that generated the best list on shopping cart?
    - What is the algorithm that generated the worst list on shopping cart?
    - Is there a difference on users' perception when products list contains mixed algorithms results on shopping cart?

As we were interested in comparing mean responses, data were analyzed using two statistical tests: the *t test* and the *Tukey test*. Both tests were applied to verify whether the mean responses were different, when compared between two algorithms or two different moments. In both tests, when $p > 0.05$, the means were considered different.

The R programming language text mining procedures (*tm* package) were used to analyze the open-ended question. The most frequent terms were identified in the answers given for the product selection moment and for the shopping cart moment. Answers texts were processed as follows:

1. All words were set to lowercase. Numbers, punctuation marks and stop-words were removed;
2. Pre-processed words were reduced to radicals;
3. Items frequency was calculated considering words radicals;
4. The most frequent words were identified to analyze users' perceptions of recommendations in different shopping steps.

Quantitative and qualitative data were then combined to test the hypotheses of this study and to identify Internet shoppers preferences for Top-N product recommendation regarding the shopping steps.

### C. Threats to Validity

Construct validity is the one that regards assuring the survey instrument measures what is intended to. To reduce bias in this aspect we have used a real database, with real images for respondents' evaluation. We also combined quantitative and qualitative result to confirm measurements.

Internal validity was threatened by the possibility that users pointed out preferences that were not related to the algorithm result, but to other issues. We attempted to reduce this threat by repeating the evaluation with three different products and changing the order recommended products were shown. We wanted respondents to focus on the given *recommendation* and not on the *product* itself.

Regarding external validity, our results cannot be generalized to other contexts. We have received an expressive number of responses, however our results are still prone to validation in other contexts.

## IV. Results

From 261 people that started answering to our research, we got 202 complete answers to the survey, from which there are 38% female and 62% male respondents. Their ages range from 18 to 67 years-old, with 28 years-old on average. Most of the respondents have experience with Internet shopping: 44% of them used an e-commerce more than six times in the last six months, 34% used from three to six times, and 22% had bought from no to two times in the last six months. To the open-ended question, we got 56 responses for the product selection moment and 41 responses to the shopping cart moment.

As explained in Section III we presented users four different lists of recommended products. One of them was based on a clustering algorithm (CB), the other was based on an association rules algorithm (CF), the other list presented first products based on clustering and next products based on association rules (CB+CF), and the last list presented first products based on association rules and next products based on the clustering algorithm (CF+CB).

We got from respondents, for each list, a ranking that ranged from 1 to 5, meaning totally dislike to totally like this recommendation order. We also obtained responses regarding the product selection moment and the shopping cart moment, to identify whether there are differences on the preference at different shopping moments.

This section presents our results. We show, first, data that support the verification of hypothesis 1, then data for hypothesis 2 and, last, our open-ended question analysis.

### A. General Evaluation

According to our respondents, the worst mean was given to the CF algorithm at the product selection moment. It means that respondents dislike seeing related products (through past sales) in the moment they are analyzing their choice. Table I shows all algorithms means and the standard deviation.

The best products recommendation was generated by the CB algorithm in the moment of product selection. Next, respondents liked the hybrid list CB+CF, also in the product selection moment. The third best classification was given to the CF generated-list, in the shopping cart moment.

### B. Preferences for Product Selection Moment

To verify whether people prefer to see first similar items on product selection, we compared the mean responses given for the algorithms in this shopping step. According to our respondents, the mean rating given to algorithms is different according to both statistical tests, with the exception of the comparison between CF+CB and CF, in which the mean rating appeared to be similar to the Tukey test. Nevertheless, the p value was close to 0.05. Table II shows the comparisons between algorithms.

Analyzing the mean values, we observe that the algorithm that shows similar items (CB) is the preferred one (mean rating 3.4637) when compared with CF (2.5165), CF+CB (2.7310) and CB+CF (2.9785). We also see the preference for seeing first similar items because the mean for CB+CF (2.9785) was greater than for CF (2.5165) and CF+CB (2.7310), which were the algorithms that showed first the products related through sales (and not similar ones).

TABLE I. Mean ranking and standard deviation for each algorithm.

| Moment | Algorithm | Mean | Std. Dev. |
|---|---|---|---|
| Product Selection | CB | 3.4637 | 0.8594 |
| Product Selection | CB+CF | 2.9786 | 0.8158 |
| Shopping Cart | CF | 2.9488 | 1.0329 |
| Product Selection | CF+CB | 2.7310 | 0.8416 |
| Shopping Cart | CB | 2.7277 | 1.1504 |
| Shopping Cart | CB+CF | 2.6683 | 0.9553 |
| Shopping Cart | CF+CB | 2.6485 | 0.8601 |
| Product Selection | CF | 2.5165 | 0.9671 |

### C. Preferences for Shopping Cart Moment

We identified that hybrid approaches, when compared with CB ones, do not trigger particular preferences in users in the

moment of shopping cart. Both statistical tests agreed when means were different and similar, as shown in Table III. There was no significant difference in the rating given by respondents in the comparison of CB (mean rating 2.7277) and CB+CF (2.6683), of CB+CF (2.6683) and CF+CB (2.6485), and of CB (2.7277) and CF+CB (2.6485).

TABLE II. COMPARISON OF ALGORITHMS IN **PRODUCT SELECTION** MOMENT.

| Algorithm 1 | Algorithm 2 | t Test Res. (p) | Tukey Test Res. (p) | Mean 1 | Mean 2 |
|---|---|---|---|---|---|
| CB | CF | Different (0.0000) | Different (0.00000) | 3.4637 | 2.5165 |
| CB | CF+CB | Different (0.0000) | Different (0.00000) | 3.4637 | 2.7310 |
| CB | CB+CF | Different (0.0000) | Different (0.00000) | 3.4637 | 2.9785 |
| CB+CF | CF | Different (0.0000) | Different (0.00000) | 2.9785 | 2.5165 |
| CB+CF | CF+CB | Different (0.0028) | Different (0.0112) | 2.9785 | 2.7310 |
| CF+CB | CF | Different (0.0178) | Similar (0.0504) | 2.7310 | 2.5165 |

However, when considering the comparison of CB (2.7277) and hybrid approaches with CF, there are significant differences. Table III shows that the CF-generated list is the preferred one for the three comparisons (mean rating 2.9488). This result shows that, in the moment of shopping cart, users do prefer to see products related in past sales, that is, the variety (and not similarity) of products is valued at this moment.

### *D. Comparing Shopping Moments*

When a specific algorithm is used in different shopping moments, users perception is also different. The greatest mean was given in the product selection moment (3.4637), which means that users prefer a CB-generated list when products are being viewed. The same conclusion applies when the list showed first CB-generated products (2.9785), and then CF-generated products. Users do prefer to see first similar products. Table IV shows that CB algorithm means were different in the product selection compared with the shopping cart.

TABLE III. COMPARISON OF ALGORITHMS IN **SHOPPING CART** MOMENT.

| Algorithm 1 | Algorithm 2 | t Test Res. (p) | Tukey Test Res. (p) | Mean 1 | Mean 2 |
|---|---|---|---|---|---|
| CF+CB | CF | Different (0.0016) | Different (0.0006) | 2.6485 | 2.9488 |
| CB+CF | CF | Different (0.0048) | Different (0.0019) | 2.6683 | 2.9488 |
| CB | CF | Different (0.0427) | Different (0.0382) | 2.7277 | 2.9488 |
| CB | CB+CF | Similar (0.5726) | Similar (0.9909) | 2.7277 | 2.6683 |
| CB+CF | CF+CB | Similar (0.8268) | Similar (1.0000) | 2.6683 | 2.6485 |
| CB | CF+CB | Similar (0.4337) | Similar (0.9530) | 2.7277 | 2.6485 |

On the other hand, the CF algorithm got the greatest mean in the shopping cart moment (2.9488). When compared with the product selection moment, the CF-generated list is preferred in the shopping cart. The same behavior was observed when the CF-generated list at the shopping cart (2.9488) is compared with the CF+CB in the product selection (2.7310). The first option is the preferred one.

### *E. Comparing Algorithms*

When comparing each algorithm with all of the others, considering both shopping moments, most users ratings differed when comparing one algorithm at the moment of product selection with another algorithm at the moment of the shopping cart. Table V shows the details.

A general analysis shows that clustering algorithm (CB) is the preferred one on product selection either when used as the only algorithm (mean rating 3.4637) or combined with association rules (2.9785).

When the products presented at the product selection moment were generated by the association rules (CF) algorithm (mean ratings 2.5165 and 2.731), for most comparisons, statistical tests did not identify differences in the rated means. The only situations when there was a difference was when compared with the results of association rules algorithm at the shopping cart moment (mean rating 2.9488), in which it was better at the shopping cart moment in both cases.

### *F. Open-ended Questions Results*

We also conducted an automatic text analysis in the answers regarding respondents' opinion about the received recommendations. Figure 1 shows the bigram frequencies for the answers in the product selection moment and Figure 2 shows the frequencies for the answers in the shopping cart moment. We show bigrams with three or more occurrences. Some bigrams have three terms because of translation issues from Portuguese language.

Although term frequency metrics do not allow a semantic analysis, we can observe some interesting patterns. For the product selection moment (Figure 1) the bigram "similar products" is the one that appears the most, with ten occurrences. It might be an evidence that users do prefer to see similar products in this step of shopping process.

For the shopping cart moment, Figure 2 shows that the bigram that appears the most is "complementary products", with eight occurrences. For us, it is an evidence of the emphasis on seeing products that complement the ones one has in the shopping cart.

These qualitative results confirm, thus, the perception we got from quantitative analysis, which is the preference users have for similar products in the product selection moment and different products in the shopping cart moment.

## V. DISCUSSION

Customers in e-commerce environments may not consider algorithm accuracy the most important aspect of their

experience of recommendation systems [4]. As researchers move from open, public databases as field of study to real users perceptions, other challenges are posed and results may be surprising [4, 20].

TABLE IV. COMPARISON OF THE **PERCEPTIONS FOR EACH ALGORITHM** IN PRODUCT SELECTION (PS) AND SHOPPING CART (SC).

| Algorithm PS | Algorithm SC | t Test Res. (p) | Tukey Test Res. (p) | Mean 1 | Mean 2 |
|---|---|---|---|---|---|
| CB | CB | Different (0.0000) | Different (0.00000) | 3.4637 | 2.7277 |
| CB+CF | CB+CF | Different (0.0004) | Different (0.0003) | 2.9785 | 2.6683 |
| CF | CF | Different (0.0000) | Different (0.0000) | 2.5165 | 2.9488 |
| CF+CB | CF+CB | Similar (0.3304) | Similar (0.9416) | 2.7310 | 2.6485 |

TABLE V. COMPARISON OF MEANS FOR **ALL ALGORITHMS** FOR PRODUCT SELECTION (PS) AND SHOPPING CART (SC).

| Algorithm PS | Algorithm SC | t Test Res. (p) | Tukey Test Res. (p) | Mean 1 | Mean 2 |
|---|---|---|---|---|---|
| CB | CB | Different (0.0000) | Different (0.0000) | 3.4637 | 2.7277 |
| CB | CF | Different (0.0000) | Different (0.00000) | 3.4637 | 2.9488 |
| CB | CB+CF | Different (0.0000) | Different (0.0000) | 3.4637 | 2.6683 |
| CB | CF+CB | Different (0.0000) | Different (0.0000) | 3.4637 | 2.6485 |
| CB+CF | CB | Different (0.0118) | Different (0.0095) | 2.9785 | 2.7277 |
| CB+CF | CF | Similar (0.7486) | Similar (0.9998) | 2.9785 | 2.9488 |
| CB+CF | CB+CF | Different (0.0004) | Different (0.0003) | 2.9785 | 2.6683 |
| CB+CF | CF+CB | Different (0.0000) | Different (0.0000) | 2.9785 | 2.6485 |
| CF | CB | Similar (0.0464) | Similar (0.0576) | 2.5165 | 2.7277 |
| CF | CF | Different (0.0000) | Different (0.0000) | 2.5165 | 2.9488 |
| CF | CB+CF | Similar (0.1132) | Similar (0.3866) | 2.5165 | 2.6683 |
| CF | CF+CB | Similar (0.1479) | Similar (0.5757) | 2.5165 | 2.6485 |
| CF+CB | CB | Similar (0.9738) | Similar (1.0000) | 2.731 | 2.7277 |
| CF+CB | CF | Different (0.0206) | Different (0.04392) | 2.731 | 2.9488 |
| CF+CB | CB+CF | Similar (0.4843) | Similar (0.9873) | 2.731 | 2.6683 |
| CF+CB | CF+CB | Similar (0.3304) | Similar (0.9416) | 2.731 | 2.6485 |

This study aimed at contributing to fill the gap of users experience in e-commerce products recommendation by answering the question on *"What are Internet shoppers preferences for Top-N product recommendation regarding the shopping steps?"*. We built recommendations solutions with a real furniture web retailer and conducted a survey to identify whether respondents had preferences regarding algorithms results in different shopping moments.

Our first hypothesis was that, on selecting a product, people do prefer to see first similar products. We confirmed this hypothesis. The means of the responses received showed that the content-based algorithm generated the best solution for users in the moment they are selecting the product they would like to buy. The worst result, in this context, was generated by the collaborative-filtering solution. Hybrid approaches were best rated in the moment of product selection when similar products were shown first. We also confirmed that results generated by a single algorithm do not necessary trigger a better perception on users.

Regarding our second hypothesis, we verified the same questions, but with emphasis in the moment of the shopping cart. The hypothesis stated that people do prefer to see first different products when viewing the shopping cart. We also confirmed this hypothesis. Our data show that, at this shopping moment, the collaborative-filtering solution triggered the best perception in users. The hybrid approaches received lower ratings and, sometimes, we could not even identify a preference in user responses.

While many studies rely on training datasets to show their algorithms accuracy and error metrics for recommendation systems [3], others have shown that variety and diversity are also important when recommending items to users.

Most research focus on improving algorithms accuracy assuming that "better algorithms lead to perceivably better recommendations" [4, p. 443]. However, several researchers have been arguing that user experience is beyond algorithms accuracy and other aspects must be considered, such as diversification, product expertise or privacy concerns, among others [4].

In fact, most algorithms are developed with public movie rating databases and evaluated through training sets. However, past experiences, such as the Netflix prize, have shown that academic solutions hardly fit seamlessly in industry settings. The study of [20] has shown that simpler and faster algorithms are the preferred in web retailers, in comparison with more elaborated techniques.

We have identified other studies, like ours, that concern asking users their perception of the given recommendations. The work by [9], for example, propose a top-N recommendation algorithm evaluated by real users. Their algorithm combines collaborative filtering and content-based approaches. They evaluated their proposal with an experiment with real users and could identify how their proposal increases the quality of recommendation by measuring users' satisfaction with the recommendation.

Another aspect was evaluated by [12]. They used a satisfaction measure to show that, in group recommendations, the order in which products are shown have a positive effect on recommendation. The proposal by [10] was also evaluated by users. They propose an algorithm that combines cycling and serpentining to provide to user fresher and high-quality recommendations. They evaluated user experience by investigating user activities and self-reported perceptions.

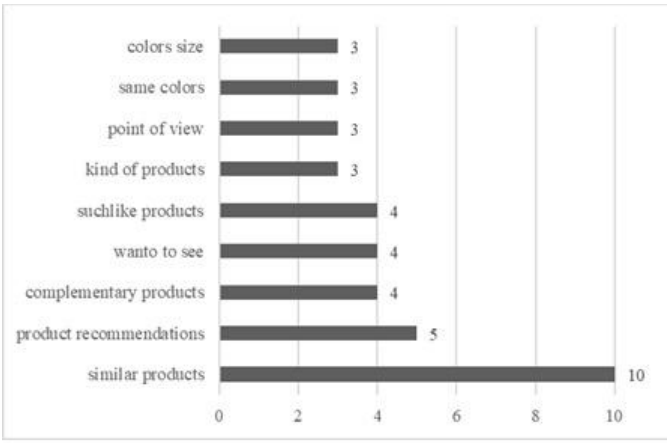


Figure 1. Bigram frequency for answers regarding product selection.

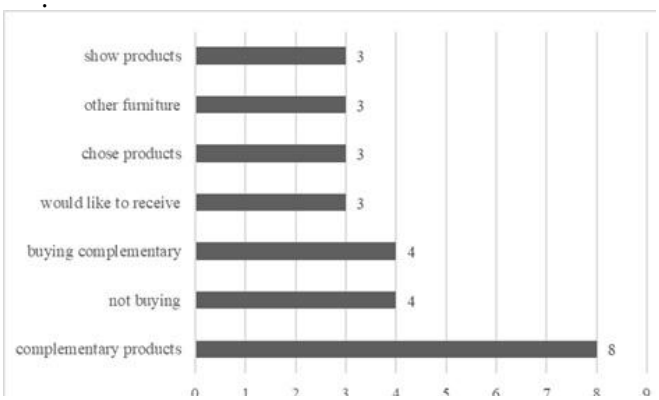


Figure 2. Bigram frequency for answers regarding shopping cart.

The work by [11] also researched user experience on recommendation systems by conducting an experiment that evaluated how different algorithms trigger different user perceptions. They surprisingly found out that the comparison of a more elaborated algorithms with a simpler one, without computational intelligence, did not create different perceptions of satisfaction on users, according to the findings by [20].

These studies confirm the evaluation of algorithms accuracy should be complemented by a user-centric perception [4]. This fact should be specially considered in e-commerce websites, where recommender systems are essential to users experience and to increase revenue [2]. Besides, e-commerce environments pose specific challenges. First because rating products is not common in these context, second because these databases are not usually made available to researchers, and lastly because these datasets are usually sparce [20]. In this context, this research contributes to the field by evaluating users' perceptions for e-commerce recommendations, using a real database from a furniture web retailer.

While our algorithmic implementations did not have a concern on accuracy or error metrics, we showed that even with simple implementations user experience is affected, exactly as shown by [20]. We consider these kinds of solutions are the ones that are viable and should be made available for small web retailers [21].

## VI. Conclusion

This study has answered the question on what the Internet shoppers' preferences are for top-N recommendations regarding different shopping steps. We implemented four different ways to recommend products and conducted a survey to ask users how they liked the recommendations received. We identified that content-based algorithms results are the preferred ones when users are selecting a product, and collaborative-filtering algorithms results are the preferred ones when users are in the shopping cart moment. These differences on users' perceptions should be considered when creating and evaluating new algorithms.

Products and sales data was based on a real Brazilian web retailer, thus our results are limited to this context. Nevertheless, our results give foundation for further research in other countries and other ecommerce environments to confirm or refuse our findings. We also contribute for other web retailers interested in improving their systems recommendations and, as a consequence, their possibility of increasing revenue.

## Acknowledgments

We thank LojasKD.com.br for providing us their sales and products database to do this research. We also thank all respondents, mainly those in the pilot test, who helped us improving our survey questions.

## References

[1] Isinkaye, F. O., Folajimi, Y. O., and Ojokoh, B. A. (2015). Recommendation systems : Principles , methods and evaluation. Egyptian Informatics Journal, pages 261–273.

[2] Usmani, Z. A., Machekar, S., Malin, T., and Mir, A. (2017). A predictive approach for improving the sales of products in e-commerce. In Proceedings of the 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics..

[3] Cremonesi, P., Koren, Y., and Turin, R. (2010). Performance of recommender algorithms on top-n recommendation tasks. In Proceedings of the RecSys 2010, pages 26–30..

[4] Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., and Newell, C. (2012). Explaining the user experience of recommender systems. pages 441–504.

[5] Sismeiro, C. and Bucklin, R. (2004). Modeling purchase behavior at an e-commerce web site: A taskcompletion approach. Journal of Marketing Research, XVL:306–323.

[6] Lin, Z. (2014). An empirical investigation of user and system recommendations in e-commerce. Decision Support Systems, 68:111–124.

[7] Li, S. S. and Karahanna, E. (2015). On line Recommendation Systems in a B2C E-Commerce Context : A Review and Future Directions . Journal of the Association for Information Systems, 16(2):72–107.

[8] Hurley, N. and Zhang, M. (2011). Novelty and diversity in top-n recommendation analysis and evaluation. ACM Transactions on Internet Technology, 10(4).

[9] Kasssák, O., Kompan, M., and Bieliková, M. (2016). Personalized hybrid recommendation for group of users:Top-n multimedia recommender. Information Processing and Management, 52:459–477.

[10] Zhao, Q., Adomavicius, G., Harper, F. M., Willemsen, M., and Jonstan, J. A. (2017). Toward better interactions in recommender systems: Cycling and serpentining approaches for top-n item lists. In Proceedings of the CSCW '17.

[11] Wojciechowski, J., Wandresen, R., Fontana, R., Marynowski, J., and Kutzke, A. (2017). Is Products Recommendation Good? An Experiment on User Satisfaction. In Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017), pages 713–720.

[12] Argawal, A., Chakraborty, M., and Chowdary, C. R. (2017). Does order matter? Effect of order in group recommendation. Expert Systems with Applications, 82:115–127.

[13] Forza, C. (2002). Survey research in operations management: a process-based perspective. International Journal of Operations & Production Management, 22(2):152–194.

[14] Pfleeger, S. L. and Kitchenham, B. (2002). Principles of Survey Research - Part 5: Populations and Samples. Software Engineering Notes, 27(5):17–20.

[15] Arthur, D. and Vassilvitskii, S. (2007). k-means++: the advantages of carefull seeding. In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, pages 1027–103.

[16] Frank, E., Hall, M. A., and Witten, I. H. (2016). The WEKA Workbench. Online Appendix for 'Data Mining: Practical Machine Learning Tools and Techniques'. Morgan Kaufmann, 4th edition.

[17] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In 20th International Conference on Very Large Data Bases, pages 478–499. Morgan Kaufmann, Los Altos, CA.

[18] Liu, B., Hsu, W., and Ma, Y. (1998). Integrating classification and association rule mining. In Fourth International Conference on Knowledge Discovery and Data Mining, pages 80–86. AAAI Press.

[19] Bodon, F. (2003). A fast apriori implementation. In Goethals, B. and Zaki, M. J., editors, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03), volume 90 of CEUR Workshop Proceedings, Melbourne, Florida, USA.

[20] Paraschakis, D., Nilsson, B., and Hollander, J. (2015). Comparative evaluation of top-n recommenders in ecommerce: an industrial perspective. In Proceedings of the IEEE 14th International Conference on Machine Learning and Applications, pages 1024–1031.

[21] Kaminskas, M., Bridge, D., Foping, F., and Roche, D. (2015). Product recommendation for small-scale retailers. Lecture Notes in Business Information Processing, 239:17–29