A Large-Scale Study of Online Shopping Behavior *

Soroosh Nalchigar
University of Toronto
Toronto, Canada
soroosh@cs.toronto.edu

Ingmar Weber
Qatar Computing Research Institute
Doha, Qatar
iweber@qf.org.qa

Parisa Lak, Ayse Bener
Ryerson University
Toronto, Canada
{parisa.lak, ayse.bener}@ryerson.ca

ABSTRACT

The continuous growth of e-commerce has stimulated great interest in generating theories and models for online consumer behavior. While studies on online consumer behavior are widespread, research on relating Internet browsing activities to online shopping behavior are scarce. This paper provides an exploratory analysis on the relationship between online browsing habits and consumers' pre-shopping effort, as one of the indicators of shopping behavior. The data used in this study was extracted from 88,637 users with more than half a million shopping instances from two large online retailers, Amazon and Walmart. Our findings provide insights for scholars to form hypotheses and design models or theories to explain online consumer behavior. Practitioners may also use the results of this study to make strategic decisions.

CCS Concepts

•Information systems → Web mining; •Applied computing → Online shopping;

Keywords

Online Consumer Behavior, Big Data, Data Science

1. INTRODUCTION

Many studies in the literature identify a growing need for discovering new knowledge, models and theories on online consumer behavior [6]. These models help industries to better understand their consumers needs and accordingly provide customized services. While studies on users' attitude concerning online shopping are widespread, studies to link users' Internet browsing habits and their online shopping behavior are scarce [4]. This study explores the relationship between users' browsing habit and their pre-shopping effort as one of the indicators of online shopping behavior.

On another note, most of the previous studies on consumers' shopping behavior use small and limited datasets. These datasets are typically limited to hundreds of participants from a specific group of individuals [7]. These data samples are considered small and biased samples by web standards. With the emergence of "big data" tools and the accessibility of web browsing information, it is possible to conduct empirical studies on real life consumer behavior. The analysis of these information may provide insights to find the emerging factors influencing users' online shopping behavior. These insights can be used to refine theories and design models to explain or predict consumers' online shopping behavior. In this study we analyze browsing data for 88,637 users, who bought more than half a million products from two major online stores. In particular, the following research questions are addressed in this study:

RQ-*What is the relationship between consumers' general web browsing habits and their pre-shopping effort?*

The implications of this study are twofold. First, the results may be used by scholars to form hypotheses and design models or theories to explain online consumer behavior. Second, practitioners may use the methodlogy used in this study to evaluate and analyze their consumers' behavior. The may also use the results from this exploratory evaluation to make strategic decisions based on the extracted behavior.

2. DATA DESCRIPTION

The dataset used in this study was extracted from Yahoo! Toolbar using big data tools such as Hadoop and Apache Pig. Data cleaning and pre-processing was performed to extract useful data for analysis using Python and Perl scripts. The dataset is prepared using the abstraction of *who* buys *what* and *how*. The following subsections briefly describe the three main data tables used in this study[1].

2.1 Users Data (Who?)

The Users Data table includes data from 88,637 users. For each user we calculated a set of attributes that are indicative of their Internet browsing behavior and interests. For each unique user, this table includes features such as the number of page views on social networks, multimedia, web search, E-mailing, news, blogs, etc. The complete list of variables in this data table is given in Appendix A, under users' browsing habits category.

*The majority of this work was done while the first two authors were at Yahoo! Research, Barcelona, Spain.

[1]Due to space limitation, the details of data preparations steps are not given here.

IDEAS '16 July 11-13, 2016, Montreal, QC, Canada

ACM ISBN 978-1-4503-4118-9/16/07…$15.00

DOI: http://dx.doi.org/10.1145/2938503.2938534

2.2 Shopping Data (How?)

The Shopping Data table includes 576,209 shopping instances performed by the users available in the Users Data table. For each shopping instance, we keep the unique users' identifier, the product ID (extracted from the product view URL), as well as the shopping time. We also included attributes related to pre-shopping effort in this table. In particular, we extracted the number of product page views, the number of queries issued on the shopping website, the number of views on price comparison and product review websites[2] within the same browsing session before each shopping instance. Also, the number of related queries in search engines, social networks, and multimedia websites before shopping instances were extracted as another attribute of pre-shopping effort (see Table 7). We calculated these attributes for Amazon and Walmart individually. The descriptive statistics for this data is presented in Table 1.

Shop	Measure	Effort Variables				
		PView	PSearch	PComp	PRev	RelSE
Amazon	Mean	12.78	8.55	0.37	0.05	0.49
	Variance	425.24	384.86	5.44	1.39	8.68
Walmart	Mean	9.72	5.59	0.29	0.03	0.06
	Variance	166.98	154.96	4.50	0.49	0.99

Table 1: Averages of pre-shopping variables for Amazon ($n = 388,236$) and Walmart ($n = 187,973$). See Appendix A for definitions of variables. Cases without product names are included here.

This table shows the average for each pre-shopping effort attribute for 576,209 shopping instances. An interesting finding from this table is that on average, online buyers tend to search and view products within shopping sites rather than looking for them in search engines, price comparison, or product review websites. We then calculated the pre-shopping effort as the sum of the quantile-shifted[3] normalized values of the five pre-shopping variables mentioned in Table 1. This variable is referred to as "effort" in our shopping data table. We also normalized the probability distributions of all the browsing variables in the Users Data table for each user.

2.3 Products Data (What?)

The Product Data table keeps the unique ID, name, category, and price of each product bought by any user. It includes data from 185,225 distinct products from 23 different categories: Appliances, Auto Parts, Babies & Kids, Beauty & Fragrances, Books, Cameras, Clothing, Computers, Electronics, Flowers & Gifts, Grocery & Gourmet, Health & Beauty, Home & Garden, Industrial Supplies, Jewelry & Watches, Movies & DVDs, Music, Musical Instruments, Office, Software, Sporting Goods, Toys, and Video Games. The description of this data table is presented in Table 2.

In this table the product categories are ranked according to the number of items bought in that category. This analysis was performed on a joint data table prepared from the shopping data table and product data table. The result is referred to as shopping-product table and includes 303,676 shopping instances. As illustrated, there are some differences observed in the distribution of categories between the two shops. For instance, in Amazon website, the highest rank belongs to Movies & DVDs and Books category, while in Walmart Home & Garden and Electronics are the most commonly bought categories. Also, some categories appearing in the top ranks in one are not necessarily observed as occurring as top ranks in the other. For example, for Walmart, Babies & Kids are among the top ten categories, while it is not the case for Amazon. Both Table 1 and Table 2 provide evidence that shopping instances are quite different between Amazon and Walmart in terms of both consumers' pre-shopping effort and product categories. Hence in the rest of this paper, to avoid side effects such as Simpson's Paradox[4], we analyze data of these shops separately.

Rank	Amazon		Walmart	
	Category	%	Category	%
1	Movies & DVDs	18	Home & Garden	23
2	Books	13	Electronics	10
3	Home & Garden	12	Clothing	9
4	Music	8	Computers	8
5	Computers	7	Babies & Kids	8
6	Electronics	6	Toys	6
7	Clothing	4	Sporting Goods	6
8	Health & Beauty	3	Video Games	4
9	Video Games	3	Appliances	3
10	Jewelry & Watches	3	Auto Parts	3

Table 2: Top ten categories in terms of percentage of stoppings for Amazon ($n = 206,328$) and Walmart ($n = 97,348$).

3. RESULTS

3.1 Browsing Interests

To find potential relationship between users' browsing interests and their pre-shopping effort, we calculated the Pearson correlation coefficient between all pairs of attributes and built the correlation matrix. To remove statistically insignificant correlations, we disregarded entries for which the $p-$value is greater than 0.01.

Some of the interesting results from this analysis for both shopping sites are listed. We found that for both shops, the fraction of views to multimedia pages (e.g., YouTube, Flickr) is positively correlated with effort spent before shopping ($r = 0.02$ for Walmart and $r = 0.006$ for Amazon). Results from Amazon show that being interested in news pages is negatively correlated with pre-shopping effort ($r = -0.05$). Similarly, users interested in art related topics spend less effort before shopping ($r = -0.04$). Other results suggest that usage of search engines is positively related to pre-shopping effort ($r = 0.04$) in both shops. Also web e-mail service usage is negatively correlated with the effort spent before shopping ($r = -0.04$, for both shops). Results from Walmart indicate that sports page views are negatively correlated with effort before shopping ($r = -0.03$). To summarize the main findings of the correlation: 1. Regular news reader and users interested in art related topics spend less

[2]These sites provide users with free services such as price comparisons, links to shopping sites, ratings and reviews.

[3]Each value is replaced by its percentile (e.g., a median value would be replaced by 0.5 and the maximum by 1.0).

[4]In this paradox, the patterns and trends that exist in separate groups of data are reversed when these groups of data are combined.

effort while making a purchase decision. 2. Multimedia webpages users tend to spend more effort on product view and search within the shopping sites. 3. Buyers who have relatively high usage of search engines spend more efforts before shopping, and most probably are not impulsive buyers. 4. Usage of web-based e-mail services is correlated with less effort spent before shopping.

These findings represent an initial attempt to understand how the users' browsing interests are related to their online shopping behavior. The p-values calculated for the correlations indicate that many browsing interest categories have significant correlation with pre-shopping effort. However, there is not any individual user who shows only one interest in their browsing habits, but there is always a set of interests in ones' behavior. These findings can have several practical and theoretical implications. Online shopping sites can use them to customize and adapt their webpage layouts and checkout process based on users browsing habits. For example, since frequent email users tend to spend less pre-shopping effort, taking advantage of cookie sharing mechanisms, shopping sites can adapt their interfaces for such users by streamlining the checkout process (e.g., highlighting the "checkout now" button or the total price).

These observations imply the need to examine why and how certain relationships exist between web browsing habits and pre-shopping efforts. This motivates use of different data collection and research methods including behavioral and qualitative methods. In addition, user experiments can validate the discovered relationships among the features and to examine their applicability for personalized user interfaces and online ads. For example, our findings suggest that using multimedia sites such as YouTube is positively related to pre-shopping effort. In other words, it seems that regular users of multimedia sites tend to spend more time on browsing and searching products before shopping. Further studies can hypothesize and explore the effects of multimedia content usage on pre-shopping efforts and study relevant factors that moderate and influence that relationship.

3.2 Clustering Online Consumers

To further examine how user segments differ in terms of shopping behavior, we performed cluster analysis. The cluster analysis is executed based on buyers' Internet usage. Among existing clustering algorithms, we used the $k-$means algorithm because (i) it produces tighter clusters than hierarchical clustering techniques, (ii) it is simple and intuitive, and (iii) comparative studies show that when the dataset is huge (as it is in our case), it has a better performance than hierarchical [1]. To choose an appropriate number of clusters of buyers for each shop, we implemented the so-called "Elbow method". This method resulted in k=5 for Amazon and k=4 for Walmart. For each shop, we performed the $k-$means clustering method 100 times to get the best results. We kept the one with the maximum percentage of between cluster distance out of total distance.

Table 3 presents the size of clusters and shows the average effort and the median price spent by users within the clusters. Tables 4 and 5 present results about browsing behavior and the top product categories for each of the clusters.

These results indicate that in Amazon the *first cluster* mainly includes users who tend to spend a lot of time on multimedia pages, as well as reading blogs, and also have a slight tendency towards arts, games, and science. For these users, categories such as home & garden, computers, video games, and sporting goods are more popular than for overall Amazon users. The *second cluster* is formed by social network users who tend to use more the interactive features of the Internet to communicate with others. For these buyers, the fraction of views on game and shopping tend to decrease and product categories such as electronics and video games rank higher. Surprisingly, we found that although the number of views on sport pages tends to decrease in this cluster, the sporting goods category has a higher rank. The *third cluster* is mainly formed by shoppers who use Internet to navigate and browse shopping sites and related pages. Besides, they have a relatively high fraction of views on search services, and also views on business related pages. For these users, we observed that categories home & garden, health & beauty, and jewelery & watch have higher ranks. It should be noted that in comparison to other clusters, these users tend to spend higher amount of efforts before shopping. The *forth cluster* of Amazon is formed by buyers whose Internet browsing include more health, home, and society related webpages. The ranking of top ten product categories for these users seems to be similar to the cluster independent ranking, except for the jewelery & watch category which is higher and for health & beauty as well as video games which are lower. The *fifth cluster* seems to be formed by users who have various interests and use Internet for different purposes, among others for e-mailing, news reading, and adult-content views. For these users, jewelery & watch, and sporting goods categories have higher rank. Users from this cluster spent the least shopping effort (on average) and tend to buy cheaper products (based on median prices) than shoppers from other clusters (see Table 3).

Shop	Cluster	Size	Avg. Efforts	Med. Price
Amazon	1	6,432	1.408	35.99
	2	20,525	1.405	30.98
	3	13,522	1.449	31.39
	4	15,378	1.401	31.00
	5	7,784	1.295	30.00
Walmart	1	14,566	1.071	69.00
	2	9,212	1.106	66.31
	3	4,911	1.054	71.91
	4	5,546	1.089	68.80

Table 3: Size, average effort spent and median prices for different clusters.

For Walmart, we found similar clusters to Amazon, indicating that the general browsing behavior does not depend too much on which online shop a user visits. The *first cluster* is formed by social networkers. Categories such as babies & kids, and auto parts have a higher rank for these users. The *second cluster* of Walmart shoppers includes general users, who use Internet for various purposes, e.g., home, health, and science. For these users, clothes have a higher rank and electronics has a lower rank. Buyers from this category, on average, tend to spend the highest shopping effort and pay the lowest price (based on median prices within the cluster). The *third cluster* includes mostly multimedia and e-mail users. For these users, sporting goods have higher rank and clothing and toys have lower ranks. Buyers from these cluster, on average, have paid the highest prices (based on median prices within the cluster) and spent the least effort, similar to the first and last clusters of Amazon. The *forth cluster* includes web searchers and shoppers. For these buy-

Shop	Cluster	Browsing habits of centroids (%)
Amazon	1	Multimedia (0.45^{+}), Arts (0.04^{+}), Blogs (0.01^{+}), Games (0.01^{+}), Science (0.006^{+})
	2	Social Networks (0.79^{+}), Shopping (0.08^{-}), Games (0.008^{-}), Sports (0.007^{-})
	3	Shopping (0.44^{+}), Search (0.25^{+}), Business (0.09^{+}), News (0.06^{+}), Home (0.03^{+}), Society (0.03^{+}), Recreation (0.02^{+}), Reference (0.02^{+}), Science (0.008^{+}), Health (0.007^{+})
	4	Society (0.03^{+}),Home (0.02^{+}),Recreation (0.02^{+}),Games (0.01^{+}),Science (0.006^{+}),Health (0.005^{+})
	5	E-mailing (0.25^{+}), News (0.2^{+}), Social Networks (0.11^{-}), Adult-content (0.10^{+}), Business (0.09^{+}), Arts (0.05^{+}), Home (0.03^{+}), Society (0.03^{+}), Recreation (0.02^{+}), Sports (0.02^{+}) Games (0.01^{+}), Science (0.007^{+}), Health (0.005^{+})
Walmart	1	Social Networks (0.83^{+}),Shopping (0.07^{-}),Business (0.03^{-}),Adult-content (0.009^{+}),Home (0.008^{-})
	2	Home (0.021^{+}), Adult-content (0.01^{+}), Games (0.01^{+}), Science (0.005^{+}), Health (0.004^{+})
	3	Multimedia (0.21^{+}), Social Networks (0.17^{-}), E-mailing (0.14^{+}), Adult-content (0.08^{+}), News (0.08^{+}), Arts (0.04^{+}), Home (0.022^{+}), Sports (0.01^{+}), Science (0.005^{+})
	4	Search (0.66^{+}), Shopping (0.26^{+}), Social Networks (0.12^{-}), Business (0.1^{+}), News (0.06^{+}), Home (0.031^{+}), Society (0.03^{+}), Recreation (0.02^{+}), Reference (0.02^{+}), Adult-content (0.01^{+}), Science (0.005^{+})

Table 4: Online consumer clusters and their browsing habits. The $^{+}$ and $^{-}$ indicate that a certain feature is over- or under-expressed compared to the corresponding cluster-independent first and third quartile of the feature. For example, Multimedia (0.45^{+}) means that the average of (normalized) total number views on multimedia sites for members of the corresponding cluster is 0.45 and this is more than the third quartile of the variable total number views on multimedia sites for all consumers together, regardless of clusters.

Shop	Cluster	Ranking of product categories
Amazon	1	Home & Garden (2,↑), Books (3,↓), Computers (4,↑), Music (5,↓), Video Games (7,↑), Clothing (8,↓), Sporting Goods (9,↑)
	2	Electronics (5,↑), Computers (6,↓), Video Games (7,↑), Clothing (8,↓), Health & Beauty (9,↓), Sporting Goods (10,↑)
	3	Home & Garden (2,↑), Books (3,↓), Health & Beauty (7,↑), Clothing (8,↓), Jewelry & Watches (9,↑), Sporting Goods (10,↑)
	4	Jewelry & Watches (8,↑), Health & Beauty (9,↓), Video Games (10,↓)
	5	Jewelry & Watches (9, ↑), Sporting Goods (10, ↑)
Walmart	1	Babies & Kids (2,↑), Electronics (3,↓), Clothings (4,↓), Computers (5,↓), Auto Parts (9,↑), Appliances (10↓)
	2	Clothing (2,↑), Electronics (3,↓)
	3	Computers (3,↑), Clothing (4,↓), Sporting Goods (6,↑), Toys (7,↓)
	4	Appliances (8,↑), Video Games (9,↓), Health & Beauty (10,↑)

Table 5: Online consumer clusters and product categories. The ↑ and ↓ show the changes in the ranking of the top ten categories bought within each cluster compared to a cluster-independent category ranking. For example, regarding the first cluster of Amazon, Home & Garden (2,↑) means that this category is the second ranked product category bought by this cluster and its ranking is higher than its ranking in the Table 2.

ers, categories of appliance and health & beauty have higher rank and video games has a lower rank.

Further, we investigated the difference between consumers' pre-shopping effort on different clusters by employing Kruskal-Wallis analysis since the samples are independent from each other and they are not related. Results of Amazon indicate that pre-shopping effort differs significantly across the five clusters ($\chi^2(4) = 613.44$, $p < 2 \times 10^{-16}$). Similarly, the result from Kruskal-Wallis analysis on Walmart clusters indicate that pre-shopping effort differs significantly across the four clusters, ($\chi^2(3) = 46.71$, $p = 3.998 \times 10^{-10}$). This result shows that users with similar browsing interest have a similar behavior in terms of pre-shopping effort.

The result is aligned with [2] where the authors segmented online shoppers based on Internet usage into three clusters. Clusters 3 in Amazon and 4 in Walmart from our study are similar to their cluster "lurking shoppers" such that the clusters mainly include consumers who use the Internet to navigate and to heavily shop. Clusters 2 in Amazon and 1 in Walmart are similar to "social thrivers" in that study, since the clusters include consumers who exploit more interactive features of the Internet for social interactions. Cluster 5 in Amazon and 3 in Walmart match with the cluster "basic communicators" in that study. These include consumers who use the Internet mainly to communicate via e-mail. It should be noted that in comparison to that work, our study was based on large-scale data and resulted in more detailed clusters. Although the clusters within two shops seems very similar, we do not expect their shopping categories to be similar, given the differences in the two shops.

3.3 Customer Loyalty

Success of online shops depends largely on customer satisfaction and other factors that will eventually increase customers' loyalty [5]. In this section, we investigated pre-shopping effort within various levels of loyalty. To measure customer loyalty, we use one of the early, and widely used definitions that is *repeated purchasing* [3]. In particular, we investigate how pre-shopping efforts change as customers get more experience with the shopping websites due to the increase in their loyalty. In this study, we count and accumulate the "loyalty level" of a shopping instance for a consumer on Amazon or Walmart. A level 1 corresponds to the first time the user buys and item on that shop during our 13 months window. A level 2 indicates the second time and so on. Using this data, we plot the effort for different loyalty levels and examine the trend. Figures 1 and 2 show the results for Amazon and Walmart respectively. In these figures, each data point is the average of the effort spent by buyers for the corresponding loyalty level and includes

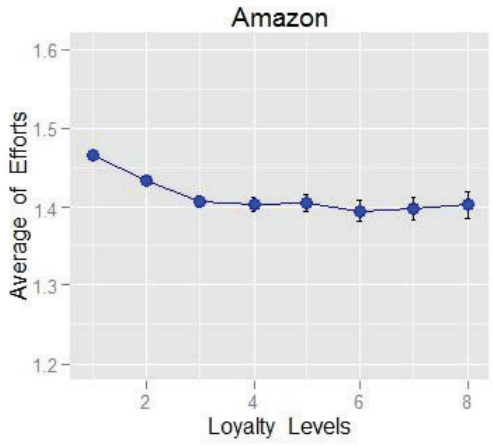

Figure 1: Shopping effort drops with an increase of loyalty to Amazon.

Figure 2: Shopping effort drops with an increase of loyalty to Walmart.

at least 400 instances in each level. Black lines are interval estimation (with 95% confidence) for the mean effort at the corresponding loyalty level.

Results indicate that an increase in the loyalty level is negatively related to the amount of effort that buyers spend before shopping. In other words, as users buy more products from a shopping site, they tend to spend less effort before the next shopping instance. To test the differences between pre-shopping efforts among eight loyalty levels, we used a one-way ANOVA. Again, the two e-tailers were analyzed individually. ANOVA assumptions were tested prior to the analysis. Q-Q plot on normality variable did not show any significant violations from normality, as expected due to the large sample size. Homogeneity of variances assumption was also tested with Bartlett test of homogeneity of variances. For Amazon and Walmart, the test resulted in ($\chi^2(7) = 2.8479$, $p = 0.8987$) and ($\chi^2(7) = 2.474$, $p = 0.929$) respectively. The p-value for both tests is greater than 0.05, meaning that we cannot reject the null hypothesis that the variance is the same for all loyalty levels.

The results from ANOVA analysis on Amazon customers

Shop	Hedonic/ Utilitarian	Category	% and direction of change
Amazon	Hedonic	Clothing	7%, Increase
		Babies & Kids	6%, Increase
		Toys	6%, Increase
		Health & Beauty	5%, Increase
		Appliances	5%, Increase
		Books	6%, Decrease
		Movies & DVDs	7%, Decrease
		Jewelry & Watches	5%, Increase
	Utilitarian	Sporting Goods	7%, Increase
		Musical Instruments	8%, Increase
		Sporting Goods	7%, Increase
		Home & Garden	5%, Increase
		Software	8%, Decrease
Walmart	Hedonic	Flowers & Gifts	32%, Decrease
		Toys	5%, Increase
		Health & Beauty	9%, Increases
		Grocery & Gourmet	19%, Increase
	Utilitarian	Computers	8%, Decrease
		Electronics	12%, Decrease
		Cameras	7%, Decrease
		Home & Garden	5% , Increase

Table 6: Significant changes in pre-shopping effort for various product categories ($p-$value < 0.01).

indicate that pre-shopping effort differs significantly across loyalty levels ($F(7, 91109) = 31.08$, $p < 2 \times 10^{-16}$). Likewise, results for Walmart consumers indicate significant difference among pre-shopping efforts across the loyalty levels ($F(7, 43187) = 6.998$, $p < 2.31 \times 10^{-8}$). The figures suggest that an increase in the consumers' loyalty comes with decreases in the pre-shopping effort to a certain level. To further investigate this threshold, we used Tukey's HSD tests and compared average efforts for each pair of loyalty levels. However, the test on our large dataset was not able to detect any threshold level. This would suggest that the means of groups are not clustered centrally, rather at least two or more means are clustered at the two extremes. One of the future directions of this research may be to find a threshold level so that different post hoc tests could be applied.

3.4 Product Type and Category

This section compares pre-shopping efforts within different product categories. We calculated the average effort for each category and compared it, using t-test, with the category-independent average of efforts. We report the results that are statistically significant ($p-$value < 0.01) and included at least a 5% relative change in the macro average. We mapped our product categories (see Section 2.3) into the two groups of *utilitarian* and *hedonic*, based on study by Kushwaha et al. [8], and investigated differences between pre-shopping effort within these two categories.

Table 6 provides the percentage and direction of changes in pre-shopping effort for different product types and categories. It shows that within both shops, the category home & garden increases the efforts by 5% and the category health & beauty increases the efforts by 5% in Amazon and 9% in Walmart. We found that the software category has an 8% lower effort in each shop. Also, in Amazon categories such as auto parts, jewelery & watches, clothing, and musical instruments increase the efforts by 5%, 5%, 7%, and 8% respectively. Categories such as books, movies & DVDs, and video games decrease effort by 6%, 7%, and 5%. For Walmart we found that categories such as beauty & fragrances

Category	Variable	Definition
Users' Browsing Habits (Who?)	UserSocialNetrwork	Number of views to social network websites (e.g., FaceBook, Twitter)
	UserAdult	Number of views to adult content websites
	UserMultimedia	Number of views to multimedia sites (e.g., YouTube, Flicker, Hulu)
	UserSearch	Number of web searches
	UserMail	Number of views to E-Mail websites
	UserNews	Number of views to news websites
	UserBlog	Number of views to blog pages
	UseArt	Number of views to webpages with art topic
	UserBusiness	Number of views to webpages with business topic
	UserGame	Number of views to webpages with game topic
	UserHealth	Number of views to webpages with health topic
	UserHome	Number of views to webpages with home topic
	UserRecreation	Number of views to webpages with recreation topic
	UserReference	Number of views to webpages with reference topic
	UserScience	Number of views to webpages with science topic
	UserShop	Number of views to webpages with shopping topic
	UserSociety	Number of views to webpages with society topic
	UserSport	Number of views to webpages with sport topic
Pre-Shopping Effort (How?)	PView	Number of product views before buying the product
	PSearch	Number of product searches before buying the product
	PComp	Number of price comparison sites visited before buying the product
	PRev	Number of product review pages visited before buying the product
	RelSE	Number of related web searches before buying the product
Product (What?)	Category	Product category (e.g., Home & Garden)
	Price	Estimated price of the product

Table 7: Research Variables

and office increase efforts by 5% and 15% and the category flowers & gifts decrease the efforts by 32%.

Results also show that pre-shopping effort differs among product types. However, when categorizing the products within the two groups of hedonic and utilitarian, no significant trend was observed. For example, within the hedonic product type in Amazon the clothing category is associated with an increase in effort, while books category (again hedonic) is associated with decrease in effort. Similar results were observed for Walmart.

4. CONCLUSIONS

User browsing data provide valuable information that can be used by both researchers and practitioners to improve their understanding of consumer behavior. The correlation analysis confirms that the information extracted from general browsing interests is commonly related with shopping behavior. The cluster analysis shows the possibility of using users' browsing behavior as the input to the design of personalized marketing and targeted advertisement. The analysis on consumers' loyalty suggests that as users get more familiar with the website they tend to spend less effort before their upcoming purchase(s). Moreover, our analysis shows that the product category can not be ignored while evaluating consumers' pre-shopping effort.

In this work, we focused on shopping instances of two online retailers, Walmart and Amazon. Although these shops are among top shopping sites, this may limit the generality of the findings. Also, we performed an in-session analysis of shopping behavior, where in practice shoppers may invest significant effort in earlier sessions, as well as outside the web. It is evident that more experimentation, with other data sources and different research methods is required, to generalize our results and further investigate online shopping behaviour. In the presence of additional data sources such as email or instant messenger, our study could be extended further to incorporate social networking information.

5. REFERENCES

[1] O. A. Abbas et al. Comparisons between data clustering algorithms. *International Arab Journal of Information Technology*, 5(3):320–325, 2008.

[2] M. Aljukhadar and S. Senecal. Segmenting the online consumer market. *Marketing Intelligence & Planning*, 29(4):421–435, 2011.

[3] G. Brown. Brand loyalty - fact or fiction? *Advertising Age*, 23:53–55, 1952.

[4] M. K. Chang, W. Cheung, and V. S. Lai. Literature derived reference models for the adoption of online shopping. *Information & Management*, 42(4):543–559, 2005.

[5] C.-M. Chiu, H.-Y. Lin, S.-Y. Sun, and M.-H. Hsu. Understanding customers' loyalty intentions towards online shopping: an integration of technology acceptance model and fairness theory. *Behaviour & Information Technology*, 28(4):347–360, 2009.

[6] A. G. Close and M. Kukar-Kinney. Beyond buying: Motivations behind consumers' online shopping cart use. *Journal of Business Research*, 63(9):986–992, 2010.

[7] W. K. Darley, C. Blankson, and D. J. Luethge. Toward an integrated framework for online consumer behavior and decision making process: A review. *Psychology & marketing*, 27(2):94–116, 2010.

[8] T. Kushwaha and V. Shankar. Are multichannel customers really more valuable? the moderating role of product category characteristics. *Journal of Marketing*, 77(4):67–85, 2013.

APPENDIX

A. VARIABLES

Table 7 summarizes and presents the complete list of variables along with their definitions.