Association for Information Systems
AIS Electronic Library (AISeL)

December 2006

Does Search Matter? Using Clickstream Data to Examine the Relationship between Online Search and Purchase Behavior

Andrew Lockwood
Wayne State University

Joni Jones
University of South Florida

Jian Zhang
Wayne State University

Follow this and additional works at: http://aisel.aisnet.org/icis2006

DOES SEARCH MATTER? USING ONLINE CLICKSTREAM DATA TO EXAMINE THE REALTIONSHIP BETWEEN ONLINE SEARCH AND PURCHASE BEHAVIOR

Economics and Information Systems

Neveen F. Awad
Wayne State University
nawad@wayne.edu

Joni L. Jones
University of South Florida
jjones@coba.usf.edu

Jian Zhang
Wayne State University
al2081@wayne.edu

Abstract

One of the most frequently used cues of the likelihood to purchase in a brick-and-mortar setting is the searching and browsing behavior of consumers. Retail sales employees are often trained to examine a consumer's searching behavior in determining their likelihood to purchase. It would seem, therefore, that such a similar correlation should occur in the online setting. Namely, consumes'r search behavior should be related to their likelihood to purchase. However, does search behavior offer any information in the online setting? It is this question that we explore in this paper. In addition, we investigate how the relationship between search and purchase behavior differs at different levels of analysis, extended search, and session-level search. We find that in the aggregate, when taking extended search into account, search breadth and depth are both positively associated with purchase behavior, suggesting that the extended search process is positively associated with purchase behavior. When we drill down to the session level, separating the sessions by person, we find that search breadth is negatively associated with amount spent online, whereas search depth is positively associated with amount spent online. These results show that a relationship exists between search behavior and purchase behavior. In addition, the results suggest that firms should examine both extended search and session-level search when assessing their consumer base.

Introduction

The ability to track consumer behavior online is considered a prime advantage that online retailing has over brick-and-mortar retailers. However, firms are still grappling with how to understand their consumers through click steam data. Initially, firms were concerned with visit volume, conjecturing that by increasing the number of visitors, firm revenue would also increase. In fact, visit measures continue to be among the most widely used and cited online metric (Demers and Lev 2001). However, visit measures clearly do not tell the full story (Moe and Fader 2004). Thus, practitioners and researchers alike are presented with an important issue, namely, firms are collecting myriads of click stream data online, but they do not know what the data is telling them about the relationship between consumer search and purchase behavior. Thus, in this study, we explore the relationship between search behavior and purchase behavior at two different levels of analysis; specifically, we examine search breadth and search depth in association with purchase behavior in the extended and session-level contexts. Prior literature has examined inter-firm search analytically (Bakos 1997), in relation to posted price (Brynjolfsson and Smith 2000), in an experimental setting (Lynch and Ariely 2000), and in relation to individual tendency to search over time (Johnson et

al. 2004). In this study, we extend this prior research by examining the relationship of online search behavior to online purchase behavior across sites. Our goal is assess whether a relationship between search and purchase behavior exists, and how that relationship appears different at different levels of analysis.

Prior Literature

The growth of online retailing has led many to believe that because the Internet offers lower search costs, consumer search increases (Bakos 1997). In the offline setting, there is extensive research that discusses consumer search, see Newman (1977) for a review. For example, consumers reportedly visit approximately three stores when purchasing a major appliance. Previous research has also provided some evidence that consumers limit their online search: Adamic and Huberman (Adamic and Huberman 2001) found that the top 1% of online sites capture 50% of all site visits; Johnson et al. (Johnson et al. 2004) found that, on average, households visit online 1.2 book sites, 1.3 CD sites, and 1.8 travel sites during a typical active month. Other studies regarding online consumer search have either examined consumer behavior at a given store over time (Moe and Fader 2004) or within a given session (Bucklin and Sismeiro 2000; Moe 2003); or they have examined consumer behavior across sites, including: rates of learning and purchasing across sites (Johnson et al. 2003), and individual visit behavior across firms (Park and Fader 2003). However, to date, there remains a gap in the literature regarding how consumer search, both across firms and within firm Web sites, affects consumer likelihood to purchase; it is this gap that we address in this paper.

A great deal of research has examined the patterns of consumer behavior online. For example, previous research has shown that consumers implicitly have a favorite firm when shopping online (Fader and Hardie 2001), that sites with fast learning curves are associated with higher rates of purchase (Johnson et al. 2003), and that shoppers search across very few sites in a given month (Johnson,et al. 2004). Similarly, a growing body of research shows that consumers face relatively high search costs when taking seemingly simple actions such as reading a screen of offers at a shopbot (Brynjolfsson, Dick et al. 2003), re-bidding at a name-your-own-price intermediary (Hann and Terwiesch 2003), and entering eBay auctions (Bajari and Hortacsu 2003). However, to date, the literature has not directly examined the relationship between search and purchase behavior online.

The Use of Online Click Stream Data

We utilize click stream data to examine the relationship between online consumer search and purchase behavior. Most click stream data exists independent of demographic information of the users in the interest of privacy. In order to work around this shortcoming, previous researchers have incorporated heterogeneity into empirical models based on click stream data through using stochastic distributions around key behavioral parameters (Moe and Fader 2004). When purchases are assumed to be a stochastic result of the current visit alone, then past visits to the same retailer have no effect on future purchasing behavior. Such an assumption does not seem reasonable; thus, it is important to consider both the potential effects of non-purchasing visits, as well as purchasing visits, over time, as we do in our paper.

Hypotheses

Howard and Sheth (Howard and Sheth 1969) began a stream of research that has examined consumer consideration set formation. When choosing to make a purchase, consumers classically begin by first screening the set of all possible online stores to a relevant set called a store consideration set (Fotheringham 1988), then they make purchase/consumption decisions from the retailers in the search set.

Because consumers have limited information-processing abilities and limited information acquisition abilities (Manrai and Andrews 1998) they limit the number of Web sites that they visit. We refer to this limited set of firms as the consumer's search set. There has been a great deal of research on the details of this process; for this paper, we are interested in how the depth and breath of the search set is associated with the consumer purchase amounts.

Prior literature has assessed breadth of search as a sum of the sources of search (Karson and Fisher 2005). In addition, consumer search has been theoretically modeled as a process in which the consumer seeks out additional information as a function of the expected benefit of that added information (Diamond 1987). As a consumer obtains more information from visiting additional online domains, the expected benefit of the new information decreases. As a result, the probability of increasing the search breadth, by visiting more sites, is likely to decrease as a function of

the number of sites that a consumer visits. Aligned with prior literature (Johnson et al. 2004), we model search breadth as the probability that an individual consumer (*i*) searches an xth site as a function of visiting the *(x-1)th* site:

$$\Pr[X_i = x_i] = \frac{(x_i - 1)\theta_i}{x_i}\Pr[X_i = x_i - 1], \ \ x_i = 2, 3, ...,$$

The parameter θ_i is an individual-specific search propensity parameter $(0 < \theta_i < 1)$. Because the maximum value of the search parameter θ_i is 1, the probability of searching an additional site is a decreasing function of x_i. Since the probability of search equation is a recursive relationship, we can work backwards through it to obtain a logarithmic distribution (Johnson, Kotz et al. 1993):

$$\Pr[X_i = x_i] = \frac{a_i \theta_i^{x_i}}{x_i}, \ \ x_i = 1, 2, ..., \text{ where } a = -\left[\ln(1 - \theta_i)\right]^{-1}$$

Based on this theory, we utilize the logarithmic model in our empirical estimate. Our theoretical construct of search breadth, therefore, is the logarithm of the number of unique online retailer domains that a consumer visits.   Our theoretical construct of search depth is developed in a similar manner, whereby the probability that a consumer visits another page within a specific domain is likely to decrease as a function of the number of pages within that site that a consumer visits. As shown above, our theoretical construct of search *depth*, therefore is the logarithm of the number of unique pages within an online retailer domain that a consumer visits. Prior literature suggests that consumers that purchase online do so from either an implicit favorite firm (Fader and Hardie 2001), or from a small set of a few sites (Johnson et al. 2004). Based on this prior search literature, along with our theory of diminishing returns of the number of domains, we expect that in a single session, an increased search breadth will be associated with a *decreased* amount of purchases. Hypothesis 1 follows:

> *Hypothesis 1*: In the session-level search analysis, search *breadth* will be negatively associated with an increased amount of purchase.

Our theoretical construct of search *depth* was developed in a similar way as the theoretical construct of search *breadth*. We expect that increased search depth is actually a signal of increased likelihood to purchase. It is likely that consumers with a higher session-level search depth have gone into areas of the site including the shopping cart, or the order fulfillment page, and thus are more likely to complete a transaction. As such, we expect search depth to be positively associated with purchase amount at the session level. Hypothesis 2 follows:

> *Hypothesis 2*: In the session-level search analysis, search *depth* will be positively associated with an increased amount of purchase.

We examine the theoretical constructs of search breadth and search depth on two different levels, on a session-level basis and over an extended-two month search period. Prior literature has generally observed a greater level of search activity from more frequent shoppers (Johnson et al. 2004) over time, which suggests that over an extended period of time, some households are heavier searchers, as well as heavier shoppers, than other households. Consequently, we expect to see a positive relationship between search *breadth* and purchase amount over an extended search analysis period of two months. Hypothesis 3 follows:

> *Hypothesis 3*: In the extended search analysis, search *breadth* will be positively associated with an increased amount of purchase.

In addition, we expect to see a positive relationship between search *depth* and purchase amount over the extended search period, as we believe that both metrics are measures of search activity, and therefore shopping activity, over an extended period of time. Hypothesis 4 follows:

> *Hypothesis 4*: In the extended search analysis, search *depth* will be positively associated with an increased amount of purchase.

Data

We study the relationship between search and purchase activity directly, through the examination of a large panel of Internet users over time. The data used in this study is Internet click stream data collected by comScore/Media Metric, Inc, a firm that records every URL visited by families that are members of its panel. At the time of the study, comScore maintained a panel of about 1.5 million online users whose Internet behavior was monitored over time. These panelists had agreed to install a small program, PCMeter, which runs continuously as a background application on the family's home computer and records their browsing and buying behavior online. The resultant data contains information regarding the sites users visit, as well as a time stamp of when the user visited each site. The U.S. panel we analyzed corresponded to approximately 100,000 households at any given point in time, recruited by Media Metrix to be representative of the U.S. population. We examine the online shopping behavior of each of these households during a two-month period from January to February of 2003. In addition to the click stream data, we had data regarding the demographics of each household in our sample, as well as all of the online transactions of the household in our sample. Prior literature has inferred purchase activity though URL descriptions; however, in our data set, we have the actual online transaction activity, and are therefore able to directly examine the relationship between search and purchase behavior in the online setting.

Methodology

To effectively examine search behavior, it is important observe consumer behavior over an extended period of time (Johnson et al. 2004). In industry, a search session is considered concluded when a user terminates the session, signaled by either 15 or 30 minutes of inactivity, depending upon the firm. However, it is arguable that in order to have a full perspective of consumer's decision period, one must examine their search behavior over a fixed period of time, which includes multiple sessions. Consumers may extend their search attempts over several sessions, aggregating the information to eventually form a purchase decision. Such extended searching behavior has been documented in the offline setting (Putsis and Srinivasan 1994).

We expect to find evidence of extended search in the online setting, and thus we define two levels of analysis in order to examine whether results differ when you explore extended search versus session-specific search. We examine extended-search over a two-month period of analysis at the household level; whereas we examine session-level search at the individual session level. Below we explain the methodology used at each level.

Extended-Search Analysis

For the extended-search analysis, purchase behavior is examined in relation to search behavior as well as household demographic information. In the full paper, we plan to examine four different dependent variables in the extended-search analysis; these variables include: 1) Number of purchases, calculated as the total number purchases per household; 2) Purchase spending, calculated as the total amount of dollars spent per household; 3) Session density, calculated as the number of household sessions that resulted in a purchase divided by the total number of household sessions; and 4) Domain density, calculated as the number of domains from which the household made a purchase divided by the total number of domains. However, due to space limitations, the results presented in the extended-search results section will be for just the first two dependent variables: number of purchases and purchase spending. In each of these models, we control for product price.

The independent variables in the extended-search analysis include search breadth, search depth, product price, and demographics information including: 1) household eldest age; 2) household size; 3) household highest education; and 4) Census region. Equation (1) is as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \varepsilon_i \qquad (1)$$

For the machine i, y_i is the purchase behavior, β_0 is the intercept, x_{i1} is the number of search sessions of the person, x_{i2} is average product price per session per person, x_{i3}, x_{i4}, x_{i5}, x_{i6} are household eldest age, household size, household highest education, and location region respectively. And ε_i is the error term.

The parameter β_1 is assumed to be affected by two dimensions of search: search breadth (average domains per session for the person), and search depth of the sessions (average views per domain per session for the person). Therefore,

$$\beta_1 = \alpha_0 + \alpha_1 z_{i1} + \alpha_2 z_{i2} \qquad (2)$$

where α_0 is the intercept, and z_1 and z_2 are search breadth and search depth, respectively.

Consequently, combing equation (1) and (2), we obtain the equation to be estimated:

$$y_i = \beta_0 + \alpha_0 x_{i1} + \alpha_1 (x_{i1} z_{i1}) + \alpha_2 (x_{i1} z_{i2}) + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \varepsilon_i \qquad (3)$$

Lastly, aligned with our theory, we transform the linear model into a log linear exponential regression model:

$$\ln y_i = \beta_0 + \alpha_0 \ln x_{i1} + \alpha_1 \ln(x_{i1} z_{i1}) + \alpha_2 \ln(x_{i1} z_{i2}) + \beta_2 \ln x_{i2} + \beta_3 \ln x_{i3} + \beta_4 \ln x_{i4} + \beta_5 \ln x_{i5} + \beta_6 \ln x_{i6} + \varepsilon_i \quad (4)$$

Session Level

The session-level purchase behavior is modeled as a function of search breadth, search depth, demographic information, and product price. The function can be expressed as follows:

$$y_{it} = \beta_0 + X'_{it} \beta_1 + Y'_i \beta_2 + Z'_{it} \beta_3 + \varepsilon_{it} \qquad (5)$$

Where y_{it} is the purchase behavior of household i during search session t. The purchase behavior is measured as two separate dependent variables: 1) the number of purchases per session; and 2) the total spending per session. In this work in progress paper, we only report the results for the second dependent variable, total spending per session. β_0 is the intercept of the function, representing some features that cannot be observed (It is assumed that these features are person-specific features. However, when the relaxing of this assumption does not change the results). X'_{it} is the transpose of the column vector which provides the information of the search session t of the person i. It consists of four items: 1) search breadth (domains per session); 2) search depth (views per domain per session); 3) search time; and 4) A dummy variable to separate out heavy searchers (household whose average search activity fell into a range one standard deviation about the mean were classified as heavy searchers). Y'_i is the transpose of the column vector which contains the demographic information of household i. In all the search sessions, Y'_i is therefore composed of the demographic variables: 1) household eldest age; 2) household size; 3) highest education level in the household; 4) census region; and 5) unobservable factors, such as the household's preference set. Finally, Z'_{it} represents the vector of the product price of the search session. In this study, we use average purchase price of the product for Z'_{it}. β_1, β_2, β_3 are the column vectors of coefficients for the variables X'_{it}, Y'_i, and Z'_{it}, respectively. ε_{it} is the error term, and is assumed to be composed of independent identically-distributed random variables (iid)[1].

For each household i, we replicate the above search session equation for another search session in the following equation (6). In this function, the household i is conducting the search session m:

---

[1] In fact, we find evidence of heteroscedasticity through the use of both White's test and the Breusch-Pagan test. We therefore also used Generalized Method of Moments (GMM) to control for the heteroscedasticity, and the results remained consistent the OLS results. In addition, even with the presence heteroscedasticity, we can still expect to obtain consistent results with some loss of efficiency.

$$y_{im} = \beta_0 + X'_{im}\beta_1 + Y'_i\beta_2 + Z'_{im}\beta_3 + \varepsilon_{im} \qquad (6)$$

Subtracting equation (5) from the equation (6), we obtain the equation (7) for estimation:

$$y_{it} - y_{im} = (X'_{im} - X'_{it})\beta_1 + (Z'_{it} - Z'_{im})\beta_3 + (\varepsilon_{it} - \varepsilon_{im}) \qquad (7)$$

The equation (3) measures the effects of the change of the search behaviors and the target products on the customer's purchase behavior. $\varepsilon_{it} - \varepsilon_{im}$ is still iid due to the fact that both ε_{it} and ε_{im} are assumed to be iid[2].

In accordance with out theory of search breadth and search depth, we transform the linear model into a log linear exponential regression model for estimation, expressed as:

$$\ln(y_{it} - y_{im}) = (\ln(X'_{im} - X'_{it}))\beta_1 + (\ln(Z'_{it} - Z'_{im}))\beta_3 + e \qquad (7)$$

Fixed effects technique assumes that differences across categories can be captured in differences in the constant term. One methodological option is to use this technique to control for differences across families in the session level analysis. However, the technique potentially requires a data set with small number of categories and large number of observations for each category. Otherwise, the degree of freedom would be greatly reduced and the estimate would be highly inefficient, or even inconsistent. In particular, in our data, there are totally 135,748 different families (represented by 135,748 different machine_id's) and 1,870,300 different search sessions. On average, there is only 13.78 search sessions for each family. Empirical analysis using a fixed effects model on this data structure is therefore not ideal.

Results

---

[2] Again, relaxing this result does not change the results.

Extended-Search Results

Table 1. Extended-Search Analysis of Number of Purchases			
Number of Observations:	104,305		
Adj. R sq.	0.9876		
F-value	771855.00 <.0001	t Value	Pr > |t|
Explanatory Variables	Estimates (Standard Deviation)	0.05	**<.0001**
Intercept	**-1.757 (0.010)****	**-175.69**	**<.0001**
Sessions per Person	**0.149 (0.002)****	**94.46**	**<.0001**
Sessions per Person × Average Search *Breadth* per Session	**0.204 (0.005)****	**39.99**	**<.0001**
Sessions per Person ×Average Search *Depth* per Session	**0.005 (0.001)***	**3.21**	**0.0013**
Average Product Price	**0.829 (0.0003)****	**2667.18**	**<.0001**
Household Eldest Age	**0.015 (0.003)****	**6.18**	**<.0001**
Household Size	**-0.013 (0.003)****	**-4.45**	**<.0001**
Census Region 2	**-0.010 (0.004)***	**-2.34**	**0.0194**
Census Region 3	**-0.011 (0.004)****	**-2.72**	**0.0066**
Census Region 4	0.0011 (0.05)	0.26	0.7971
Household Most Education 1	-0.003 (0.008)	-0.39	0.6995
Household Most Education 2	0.010 (0.008)	1.26	0.2064
Household Most Education 3	0.0034 (0.008)	0.41	0.6837
Household Most Education 4	**0.03186 (0.008)****	**4.02**	**<.0001**
Household Most Education 5	**0.03488 (0.009)****	**4.12**	**<.0001**

** 5% level of significance

Table 1 lists the estimates of the extended-search model; There are 104, 305 observations in the estimation. The adjusted R-square is 0.9876 and the F-value of 771855.00 indicates the overall estimates are significant with more than 99% confidence.   We see from the results of the extended-search model that both search depth (0.005) and search breadth (0.204) are significant and positively correlated with number of online transactions, confirming Hypotheses 3 and 4. In addition, when search breadth and depth are kept fixed, the number of sessions (0.149) is significant and positively correlated with number of online purchases; therefore just going online often is a significant signal that a household is a heavy consumer household. The comparative impact of the search breadth (0.204) is greater than the impact of the search depth (0.005) by more than a factor of four, suggesting that at the aggregate, a good signal of more purchase transactions is a greater number of sites searched. These results are confirmed in Table 2, which presents the extended-search-level analysis using purchase amount in dollars as the dependent variable, rather than number of purchases as presented in Table 1. In interpreting the coefficients presented in tables 1 and 2, one percent's change in any of the independent variable would lead to the corresponding coefficient's percentage change in the dependent variable. For example, for the search depth, a 1% increase in search depth corresponds to a .005% increase in purchase amount. All of the major coefficients are significant, which suggests that we are more than 99.99% confident that the coefficient is different from zero. The relationship between search breadth, search depth, and purchase amount remains consistent when the dependent variable is dollar amount, rather than number of purchases, lending further credibility to the extended-search level results.

Table 2. Extended-Search-Level Analysis of Purchase Amount			
Number of Observations:	104,305		
R-square	0.9951		
F-value	1972691		
	Estimates (Standard Deviation)	t Value	Pr > |t|
Intercept	**0.1053 (0.008)****	**13.56**	**<0.0001**
Sessions per Person	**0.1478 (0.001)****	**120.78**	**<0.0001**
Sessions per Person × Average Search *Breadth* per Session	**0.1950 (0.004)****	**49.21**	**<0.0001**
Sessions per Person ×Average Search *Depth* per Session	**0.0066 (0.001)****	**5.47**	**<0.0001**
Average Product Price	**1.0340 (0.0002)****	**4285.18**	**<0.0001**
Household Eldest Age	**0.0143 (0.002)****	**7.38**	**<0.0001**
Household Size	**-0.0098 (0.002)****	**-4.47**	**<0.0001**
Census Region 2	**-0.0077 (0.003)***	**-2.27**	**0.0231**
Census Region 3	**-0.0114 (0.003)****	**-3.66**	**0.0003**
Census Region 4	-0.0022 (0.004)	-0.61	0.5401
Household Most Education 1	-0.0032 (0.006)	-0.53	0.5949
Household Most Education 2	**0.0124 (0.006)****	**2.11**	**0.0350**
Household Most Education 3	0.00424 (0.006)	0.66	0.5096
Household Most Education 4	**0.0331 (0.006)****	**5.38**	**<0.0001**
Household Most Education 5	**0.0379 (0.006)****	**5.78**	**<0.0001**

** 5% level of significance

In examining the results of both Table 1 and Table 2, we see that the higher the education level of the highest educated member of the household, the greater the number of online transactions that the household makes. This result corresponds with prior literature that has reported that higher educated consumers have less apprehensions about sharing information, including credit card numbers, online (Awad and Krishnan 2006), and thus are likely more willing to spend money online.

Session-level Results

Table 3. Session-level Analysis of Purchase Amount			
Number of Observations:	881,881		
R-square	0.9987		
F-value	8.21E+07	<u>t Value</u>	<u>Pr > |t|</u>
<u>Explanatory Variables</u>	<u>Estimates (Standard Deviation)</u>	**-106.31**	**<0.0001**
Search Breadth	**-0.0129 (0.0001)**[*]	**29.2**	**<0.0001**
Search Depth	**0.0070 (0.0002)**[*]	**28.92**	**<0.0001**
(Search Depth)^2	**0.0009 (0.0000)**[*]	**77.34**	**<0.0001**
Search Duration Time	**0.0158 (0.0002)**[*]	**-11.07**	**<0.0001**
(Search Duration Time)^2	**-0.0002 (0.0000)**[*]	**67.26**	**<0.0001**
Search Breadth × Duration Time	**0.0017 (0.0000)**[*]	**8.25**	**<0.0001**
Heavy search	**0.0135 (0.0016)**[*]	**11216**	**<0.0001**
Average Price	**1.0131 (0.0000)**[*]	**23720.9**	**<.0001**

** 5% level of significance

Table 3 lists the estimates of the session-level model. There are 881,881 observations in the estimation. The R-square is 0.9987 and the F value of 8.21E+07 indicates the overall estimates are significant with more than 99% confidence. The model is set up as a logarithmic model, which provides estimates of the people's search behaviors' elasticity for their online purchase spending, and is aligned with out theory on search breadth and search depth. We control for average price of purchased goods in the session-level model to assess the spending amount in relation to the price of the good being searched for and purchase; clearly the more expensive the product, the greater the amount of purchase spending, as the results significantly confirm.

The results based on the session level analysis show that search depth is significant and positively associated with purchase amount, confirming hypothesis 2. Specifically, a 1% increase in search depth increases is associated with a 0.0079% increase in spending. Search breadth, however, is significant and negatively associated with session-level purchase amount, suggesting that the greater number of sites that a consumer searches across in given session, the lower the purchase amount that he will spend during that session, confirming hypothesis 1. Specifically we find that, for the individual session-level analysis, a 1% increase in search breadth is associated with a 0.0129% reduction in the amount of money spent during a session, keeping all other factors fixed. Another interesting result is the quadratic relationship between session duration time and amount spent. Our results show a significant positive relationship between time and purchase amount, but a significant negative relationship between duration time-squared and purchase amount, suggesting that consumers spend more money up to a certain period of time in the session, after which an increase in the time is actually associated with a decrease in the amount spent. Specifically, 1% increase in duration time is associated with a 0.0158% in purchase amount. However, a 1% increase in duration time-squares is associated with a 0.0002% decrease in purchase amount. The inclusion of the square terms of both search depth and search depth is aligned with prior literature (Moe 2003), which suggests that search is valuable up to a certain threshold. Our results confirm such theory in the case of search duration time, where after too much search duration time, the value is negative. The results also show that heavy searchers are significantly associated with greater purchase amounts.

Discussion

One of the most cited advantages of online retailing for firms is that they can capture data regarding all of their consumers' activities. However, to date, minimal research has examined the relationship between online search behavior and online purchase behavior using such click stream data. This study explores the relationship between

search behavior and purchase behavior at two different levels of analysis; specifically, we examine search breadth and search depth in association with purchase behavior in the extended-search level and session-level contexts. In the extended search context, search depth was found to be positive and significantly associated with number of purchase transactions, and total amount spent; this result suggests that, in the aggregate, search breadth is a good signal of consumers that shop often and that spend a lot of money.

On the other hand, in the session-level analysis, search breadth was found to be significant and negatively associated with amount spent in a session, suggesting that the more sites that a consumer visits in a session, the less money the consumer spends. This result may suggest that higher search breadth in a session indicates either: 1) price-sensitive consumers (Kim et al. 1999); or 2) non-directed shoppers (Moe 2003). Future research is needed to explore this distinction in greater detail. Search depth is positive and significant in the session-level analysis, indicating that an increase in the number of pages viewed within a given site is associated with an increase in the total amount of dollars spent in a given session. This result, in accordance with the quadratic relationship between duration time and total amount spent, corresponds with previous literature that suggests that online consumers tend to be more time constrained than the average consumer (Bellman et al. 1999). Thus, the online consumers that searches deeper into a domain in a given session are likely to be focused on purchasing during that session, and thus are correlated with a greater session-level spending amount.

Limitations

While click steam data offers a wealth of information, it also has its inherent limitations. Due to the nature of the click stream data, the analysis was performed at the household level, rather than at the individual consumer level. Correspondingly, we likely overstate both search breadth and search depth. Another potential limitation of click stream data is that it is difficult to discern which type of search activity is being performed (i.e., directed searches versus non-direct browsing); as such we categorize all types of searches together, which, again, may lead to an overestimation of search in the model. Future research is needed to develop techniques for further classifying online behavior into different types of search behavior. Lastly, another limitation of the study is that we do not have data regarding offline search and purchase behavior. Thus, consumers may search online and purchase offline, or may search offline and purchase online, and we are unable to capture those cross-channel effects.

Conclusion

This paper establishes a direct relationship between search behavior and purchase behavior online. In addition, we illustrates that the nature of the relationship between online search and online purchase behavior is dependent upon the level of analysis. From the perspective of the firm, these results can inform customer segmentation and strategies: consumers that search across more sites over an extended period of time may be more frequent purchasers, and may spend more money. On the other hand, consumers that search across more firms within a given session may spend less money. Lastly, consumers that search more *within* a firm's site within a given session are likely to spend more money. Thus, firms may want to consider up-selling to consumers with a longer search path in a given session, or may want to offer various purchase incentives during the session. This research underscores the value of looking at different levels of analysis, depending upon the information that a firm is seeking to find. We see this research as an initial step towards exploring the relationship between search behavior and purchase behavior online *directly* through click stream and transaction data. Our goal for future iterations of this paper is to further develop theory of the types of searches that consumers perform online, in a way that can be tested using efficient empirical methods. There is a wealth of theory to be developed and tested in the area of online search and purchase behavior; we believe this paper takes in an important first step in that direction.

References

Adamic, L. and B. A. Huberman (2001). "The web's hidden order." <u>Communication of the ACM</u> **44**(9): 55-59.

Awad, N. F. and M. S. Krishnan (2006). "The personalization privacy paradox: An empirical evaluation of information transparency and the willingness to be profiled online for personalization." <u>MIS Quarterly</u> **30**: 13-28.

Bajari, P. and A. Hortacsu (2003). "The Winner's Curse, Reserve Prices, and Endogenous Entry: Empirical Insights from eBay Auctions." RAND Journal of Economics **34**(2): 329-355.

Bakos, J. Y. (1997). "Reducing Buyer Search Costs:Implication for Electronic Marketplaces." Management Science **43**(12).

Brynjolfsson, E., A. A. Dick, et al. (2003). "Search and Product Differentiation at an Internet Shopbot." MIT Sloan Working Paper No. 4441-03.

Brynjolfsson, E. and M. Smith (2000). "Frictionless Commerce? A Comparison of Internet and Conventional Retailers." Management Science **46**(4).

Bucklin, R. and C. Sismeiro (2000). "A Model of Website Browsing Behavior Estimated on Clickstream Data." Journal of Marketing Research **40**(3): 249-267.

Demers, E. and B. Lev (2001). "A Rude Awakening: Internet Shakeout in 2000." Review of Accounting Studies **6**(2 - 3): 331 - 359.

Diamond, P. (1987). Search Theory in The New Palgraeve. A Dictionary of Economics. J. Eatwell, M. Milgate and P. K. Newman. London, Stockton Press.

Fader, P. S. and B. G. S. Hardie (2001). "Forecasting Repeat Sales at CDNOW: A Case Study." Interfaces **31**(3): 94-107.

Fotheringham, A. (1988). "Consumer Store Choice and Choice set Definition." Marketing Science **7**(3): 299-311.

Hann, I.-H. and C. Terwiesch (2003). "Measuring the Frictional Costs of Online Transactions: The Case of a Name-Your-Own-Price Channel." Management Science **49**(11): 1563-1579.

Howard, J. A. and J. N. Sheth (1969). The Theory of Buyer Behavior. New York, John Wiley.

Johnson, E. J., S. Bellman, et al. (2003). "What Makes A Web Site 'Sticky'? Cognitive Lock-In and the Power Law of Practice." Journal of Marketing **67**(2): 62-75.

Johnson, E. J., W. W. Moe, et al. (2004). "On the Depth and Dynamics of Online Search Behavior." Management Science **50**(3): 299-308.

Johnson, N. L., S. Kotz, et al. (1993). Univariate Discrete Distributions. New York, John Wiley & Sons, Inc.

Karson, E. J. and R. J. Fisher (2005). "Predicting intentions to return to the Web site: Extending the dual mediation hypothesis." Journal of Interactive Marketing **19**(3): 2-14.

Kim, B.-D., K. Srinivasan, et al. (1999). "Identifying price sensitive consumers: the relative merits of demographic vs. purchase pattern information - A Meta-Analysis of Econometric Models of Sales." Journal of Retailing **75**(2): 173-193.

Lynch, J. G. and D. Ariely (2000). "Wine Online: Search Costs Affect Competition on Price, Quality, and Distribution." Marketing Science **19**(1): 83-103.

Manrai, A. K. and R. L. Andrews (1998). "Two-stage discrete choice models for scanner panel data: An assessment of process and assumptions." European Journal of Operations **111**(20): 193-215.

Moe, W. W. (2003). "Buying, Searching, or Browsing: Differentiating between Online Shoppers Using In-Store Navigational Clickstream." Journal of Consumer Psychology **13**(3): 29-39.

Moe, W. W. and P. S. Fader (2004). "Capturing evolving visit behavior in clickstream data." Journal of Interactive Marketing **18**(1): 5-9.

Newman, J. R. (1977). Consumer external search: Amount and determinants. Consumer and Industrial Buying Behavior. North-Holland, New York., J. S. Arch Woodside.

Park, Y.-H. and P. S. Fader (2003). "Modeling Browsing Behavior at Multiple Sites." Marketing Science **23**(3): 280-303.

Putsis, W. P. J. and N. Srinivasan (1994). "Buying or Just Browsing? The Duration of Purchase Deliberation." Journal of Marketing Research **31**(August): 393-402.