ICEB 2013 Proceedings

International Conference on Electronic Business (ICEB)

Winter 12-1-2013

Predicting E-Commerce Adoption Using Data About Product Search And Supplier Search Behavior

Libo Li

Frank Goethals

Bart Baesens

PREDICTING E-COMMERCE ADOPTION USING DATA ABOUT PRODUCT SEARCH AND SUPPLIER SEARCH BEHAVIOR

Libo Li, Université Catholique de Lille, France, danie.aba@gmail.com
Frank Goethals, Université Catholique de Lille, France, f.goethals@ieseg.fr
Bart Baesens, University of Southampton, United Kingdom, Bart.Baesens@kuleuven.be

ABSTRACT

In this paper we use a semi-supervised learning model to predict whether a person thinks buying a specific product online is appropriate. As input, information is used about the channels one deems appropriate to find product information or to find suppliers. Both online and offline channel preferences are found to be valuable to predict e-commerce adoption. The practical consequence of the work is that (binary) data about a user's preferred channel for information retrieval can be helpful to estimate the probability the person is interested to buy a specific product online so that publicity for an online shop is only shown to people who actually believe buying that product online is appropriate. The predictive performance of our approach is considerably better than that reported in earlier research. Our results also show that semi-supervised learning has advantages in terms of predictive performance compared to supervised learning.

Keywords: E-commerce, channel acceptance, graph based semi-supervised learning, implicit network analytics

INTRODUCTION

Online sales are rising. According to Forrester research, online retail sales will reach $262 billion in 2013 in the US and €128 billion in Europe. While this number is increasing with about 10% every year, online sales still make up less than 10% of the total retail sales. Clearly, not all people think buying online is appropriate for all products. As a consequence, it is important to target publicity for online shops to people who are actually interested in buying a specific product online. In an attempt to do that, companies nowadays can use big amounts of data concerning consumers. For example, big phone companies are selling data they gather about their subscribers' locations and web-browsing habits. This paper investigates a way to predict whether a person is interested in buying something online or offline. More specifically, we investigate whether knowledge about channel choice in early data-gathering steps of the buying process is useful to predict whether a person will actually make the purchase online.

Marketing literature distinguishes several steps in the buying process. In this paper we investigate the relation between media used in three steps of that process: (1) the medium used to search product information and (2) to search supplier information and (3) to make the purchase in the end. When we talk about "e-commerce adoption" in this paper, we then refer to the use of online media in the purchasing process step where the purchase is actually made.

Prior research aimed at explaining (rather than predicting) e-commerce adoption by considering it as a specific case of technology acceptance with marketing elements [13]. The theory of planned behavior (TPB) has been extended to explain that the e-commerce adoption process is affected by product information seeking behavior [13]. This paper, however, did not investigate the predictive power of this information with appropriate predictive measures [14][15]. Vendor information seeking has been studied as well because vendor information is useful for choosing a channel to buy [7]. On the other hand, many firms attempt to participate in e-commerce by moving part of their business from offline to online platforms, e.g. Walmart acquired online service providers Oneops and TastyLabs. Still, many products and services are conventionally promoted and sold offline. This brings both opportunities and challenges for research in e-commerce adoption.

In practice predicting e-commerce adoption is a difficult task since there are many people browsing without actually purchasing. With growing attention in data analytics using machine learning and data mining techniques, we discuss predictive modeling with the semi-supervised learning paradigm to classify potential e-commerce adopters. Previous studies have shown that information of users' social ties is useful to predict e-commerce adoption [18][19]. That information, however, only has limited predictive power (the average AUC reported was 0,62; see below). Hence, other approaches with more predictive power are welcome. Given the attention that was given in prior research to channel choice for product information seeking and supplier information seeking, it is interesting to investigate whether such information is useful to predict e-commerce adoption. Hence we formulate these research questions:

- Is information about channel preferences for finding product information valuable to predict e-commerce adoption?
- Is information about channel preferences for finding supplier information valuable to predict e-commerce adoption?

In what follows we first discuss related literature. Subsequently we introduce the modeling details and discuss the design of the experiment to assess the predictive power of our model by contrasting different data mining models from different paradigms. Finally we discuss and summarize our work in the conclusions.

RELATED WORK

E-commerce acceptance theory

Much explanatory research has been done towards technology acceptance. The Theory of Planned Behavior (TPB) and related works [1][2][3][4] provided dimensions such as attitudes towards behavior, subjective norms and perceived behavioral control to explain the reasoned human behavior. *Attitudes towards behavior* then stand for the individual's evaluation of the performance from the outcome of the behavior. *Subjective norm* concerns the fact that the individual's perception of the behavior is influenced by social normative pressures for example from family and friends. *Perceived behavioral control* stands for the ease of conducting the behavior. A person's decision to perform certain behavior is affected by his/her own intention which relates to the mentioned dimensions. The TPB was widely discussed in information systems research to gain insight into e-commerce [11][13]. Pavlou et al [13] argued that both the intention to make some purchase and the intention of seeking information are important to understand the process of e-commerce adoption. They extended the TPB with extra components of self-efficacy and controllability that are:

- Individual judgments of a person's capabilities to perform a behavior [4]
- Individual judgments about the availability of resources and opportunities to perform the behavior [5]

as well as other concrete elements involving technical (e.g. download delay) and non-technical details (product characteristics) of e-commerce. That research shows e-commerce adoption is not a monolithic but a multidimensional activity that involves the information seeking process [13]. Other research pointed out that users' channel preferences are linked to relative advantages of the channels concerning issues such as trust, convenience and efficacy of information acquisition [7].

Predictive versus explanatory models

While most research on information systems acceptance concerns explanatory modeling, our paper concentrates on predictive modeling. *Explanatory models* are models that are 'built for the purpose of testing causal hypotheses that specify how and why certain empirical phenomena occur' [15, p.554]. Researchers then test causal hypotheses using association-type models, e.g. regression and structural models. Such modeling evaluates the goodness of fit of the model using criteria such as adjusted R^2. *Predictive modeling* (such as data mining) is 'designed for predicting new/future observations or scenarios' [15, p.555]. In predictive modeling, one uses a holdout sample where part of the sample is used to build the model and the other part is used to validate the predictive accuracy. These two types of models compensate for each other's shortcomings since predictive models are built based on the distribution of the data (and possible relations are not limited to hypothesized relationships) but are weak in terms of explanatory power. Explanatory models contain causal associations between the response and the constructs, but overestimate the model fit since it builds and evaluates the model using the same sample (i.e., there is no hold-out). Explanatory models can tell predictive model builders what variables could be considered to make predictions. A literature review of MISQ and ISR articles showed that predictive modeling is scarce in mainstream IS research [15] and that sometimes explanatory modeling was used to reach a predictive goal, what is inappropriate given the fact that no holdout is used. Besides, 'the best explanatory statistical model will almost always differ greatly from the best predictive model' [15, p.555].

Much IS research [6] [17] concerns e-commerce adoption but that is typically exploratory research. In this paper we work on a parallel path, using predictive models in IS acceptance research so as to come to a higher predictive accuracy than what can be achieved with explanatory models.

Semi-supervised learning

In the field of predictive modeling research techniques are used such as machine learning and data mining. The full data sample is randomly split into training and test sets to demonstrate the predictive power. A training set is used to build a predictive model whereas a test set is used to measure the predictive performance. The goal is to predict the value of a response variable ('label'), such as the binary outcome of whether a user is going to be an adopter/non-adopter. In practice there might be datasets with instances for which the response variable is fully known or only partially known. So-called *supervised learning* is applied to the scenario where all labels are known. *Semi-supervised learning* is the approach where the response variable is only known for part of the sample. Both labeled and unlabeled data are then used to build the predictive model [22]. The unknown label records can also be valuable for prediction in the sense that they share similarity in certain features with others that might have a known outcome. Despite the fact their true label is not known, these unlabeled samples can still contribute to prediction to some extent based on similarity among labeled and unlabeled data. For instance different people with similar interests in using certain online channels might be interested in shopping online. In practice, semi-supervised learning classification methods are preferred since less labeled data are needed. It has been used widely in engineering fields involving computer vision [16] and knowledge discovery on the web [20]. We use such approach in the hope to decently predict e-commerce adoption.

RESEARCH DESIGN

A survey has been conducted in 2012 at one European business school gathering information on channel preferences (1) to gather product information, (2) to find a supplier and (3) to make the purchase. Students and their parents were asked to reveal their perceived appropriateness of different media in the context of the procurement of different products. We refer to Table 1 for a list of the different products. The list includes a number of services as well (but for reasons of brevity we will talk about 'products' in this paper).

Table 1 Product item used in the survey

Product type		Percentage of people that intend to buy the product online
	2nd hand car	0.2298
	books	0.5134
	pubs and cafeterias	0.2127
	computers	0.5012
	fitness centers	0.2323
	camera	0.4328
	hotels	0.6993
	car rental	0.5770
	copy-service	0.2494
0	family doctor	0.0954
1	air-conditioning	0.3399
2	clothing	0.4523

The online purchase of a service should then be interpreted as the online reservation of a service. The products that were included in this research are the same as in [19], so as to enable a comparison of the results.

An ordinal scale from 1 to 5 (most applicable to not applicable at all) was used to reveal the appropriateness of different media (see Table 2 for an overview of the different media).

Table 2 Online/offline channels used in the survey

Search product information online	Search product information offline	Search supplier information online	Search supplier information offline
Google and other search engines	Magazine and books, family and friends, from sellers	Google and other search engines, online yellow pages, online auction sites	Magazine and books, family and friends and other people, yellow pages, already having a fixed supplier

For reasons of clarity we here wish to say explicitly that students did not have to fill out the questions about where product information was searched for 'products' such as pubs and cafeterias, fitness centers, hotels, copy services and family doctors. Hence, no information on the channel choice to search product information could be used in the predictive model for these products.

We utilized data from 409 out of 414 respondents who filled out the survey. Data has been cleaned and preprocessed. We removed observations which had more than 10% missing values.

Predictions were made for the appropriateness of buying online for some person for a specific product. That is, all products were treated separately. For each product, we formally define a feature space of subject *i* 's channel preferences $\Phi_i=\{S_i^{on},S_i^{off},P_i^{on},P_i^{off}\}$where $S_i^{on},S_i^{off},P_i^{on},P_i^{off}$ are channel preferences for searching supplier information online/offline and searching for product information online and offline respectively. The response variable – whether a respondent thinks it is appropriate to purchase online or not - is defined as $y=\{y_1,y_2\ldots y_l\ldots y_n\}$ where $\{y_1,y_2\ldots y_l\}$= subjects' outcome (label) that is known, and $\{y_{l+1}\ldots y_n\}$= subjects' outcome (label) that is unknown. n is then the total number of respondents in the sample.

We obtain a similarity score between all respondents by using different distance measures. Affinity networks can then be constructed on the basis of the channel preferences. For instance we obtain an affinity network for the way supplier information is searched online:

$$W_{ij}^{S\_on} = dist(S_i^{on}, S_j^{on}) \qquad (1)$$

The same holds for other types of affinity networks that are constructed from other features such as S_i^{off}, P_i^{on} and P_i^{off}. Different distance measures can be used, such as Euclidean distance or Cosine similarity. Applying raw numerical values of the channel preferences to these measures might not preserve the ordinal nature of the data. Therefore, the scale measures from the survey have been converted into ranks before calculating the instance-based similarity. Specifically, if a person A rates online channel preferences for 5 different media as [1 1 2 5 5] and person B gives a rating [2 2 3 4 4], they will essentially get the same ranking on the channel preferences [1.5 1.5 3 4.5 4.5]. Ranks are averaged whenever there is a tie. We denote rank(S_{ik}^{on}) as the rank of *k* th online medium to search supplier information for person *i*,

In practice, companies will generally not create surveys (as the one we created) to get the ordinal information we use. Rather one's channel preference can be observed and binary data is available on the use of some channel (e.g., did the site visitor come from a Google search results page?) and in the best case a ranking can be made of the use of different media (e.g., the person came once to our site from the yellow pages but many times from Google). Hence we also converted the 1-5 ratings into binary ratings so that a medium that was perceived as appropriate (score 1 or 2) received a score of 1 and a medium that was not really perceived as appropriate (scores 3 to 5) received the score 0. For instance A's preference of [1 1 2 5 5] then became [1 1 1 0 0]. We use Spearman distance [9] as our distance measure among all other choices. Using Spearman distance, we can for example create an affinity network based on similarity in the way supplier information is sought:

$$W_{ij}^{s\_on} = \sum_{k=1}^{n} (rank(s_{ik}^{on}) - rank(s_{jk}^{on}))^2 \qquad (2)$$

With the network in matrix *W*, a predictive model can be built by learning a function f which takes nodes in network *W* as input and predicts the label of the nodes as output. *f(x_i)* then is the predicted label for respondent *i* given x_i, the respondent's ranking of the appropriateness of different media. We denote all samples in the data with labeled samples $\{(x_1,y_1),(x_2,y_2)\ldots,(x_l,y_l)\}$ and $\{(x_{l+1},y_{l+1}),\ldots,(x_n,y_n)\}$ as unlabeled samples without y as the label. The x_i here is the channel preference such as [1.5 1.5 3 4.5 4.5] for person *i*. A regularization framework can be built around labeled and unlabeled data to guide the design of the function f .

$$\min_{f.f(x)} \alpha \sum_{i=1}^{l} (y_i - f(x_i))^2 + \sum_{i,j=1}^{n} w_{ij}(f(i) - f(j))^2 \qquad (3)$$

There are two minimization objectives in such framework: the function f should provide an accurate estimate on labeled samples, and smoothness between labeled and unlabeled samples meaning the samples close in the network should have similar labels. These two objects can be traded-off; thus balanced with parameter α in (3).

Gaussian Random Field function (GRF) [21] provides a sol1ution to the problem formulated above, by using a harmonic function that maps values from *x* to *y* in labeled data, and estimates the label of the unlabeled data with their neighbor data points in the network structure. More specifically, the labeled nodes are learned in the function as what they are; while for unlabeled nodes we intend to minimize the following function:

$$E(f) = \frac{1}{2} \sum_{i,j}^{n} w_{ij}(f(i) - f(j))^2 \qquad (4)$$

It is harmonic as it is twice continuous differentiable and $\Delta f = 0$. The harmonic property provides the unlabeled nodes:

$$f(j) = \frac{1}{d(i)} \sum_{i \sim j} W_{ij} f(i), forj = l+1,...l+u \qquad (5)$$

This basically means the labels of unlabeled nodes are decided by the average of their neighbor. The degree of a node is denoted as d(i).

In this paper we intend to illustrate that semi-supervised learning performs better than supervised learning when using labeled and unlabeled data. We will therefore compare the results achieved with GRF, with the predictive performance of a logistic regression (see below). Using the affinity network and label of subjects we build a predictive model and cross validate the prediction performance. For instance, for a 30% holdout, 70% of the data is used to train the model and the remaining 30% is only used to test the prediction performance. We contrast the prediction result between supervised learning and semi-supervised learning methods with different percentages of holdout.

Just as is the case in practice, where there is a huge difference between the number of people browsing and those purchasing, there will be a lot of data entries about customers which only browse without knowing the outcome whether they are interested in e-commerce or not. Semi-supervised learning is considered to be a good choice with its ability to classify labeled and unlabeled data. Logistic regression is used as the supervised learning benchmark. It takes in feature Ψ as input and predicts the adoption by estimating parameters a and b using the maximum likelihood procedure [12].

$$\ln(\frac{p(y = y_i \mid \Psi_i)}{1 - p(y = y_i \mid \Psi_i)}) = a + b\Psi \qquad (6)$$

AUC (area under the ROC curve) is used as the criterion to evaluate predictive performance. AUC can be considered as the probability that a classifier in the predictive model will rank a randomly selected positive case (someone's intent to purchase online) higher than a randomly selected negative case. [10]. We use AUC to measure the predictive accuracy because the e-commerce acceptance rate is low for some of the products (see Table 1). In such cases it is considered to be a better measure than accuracy for example [10]. All the tests have been repeated with 100 trials. We check the statistical significance using the Mann–Whitney U test [8] with significance level 0.95.

RESULTS

Tables 3 to 6 show the predictive performance of GRF and logistic regression with a 30% holdout (Tables 3 and 4) and 90% holdout (Tables 5 and 6) using the ranked data. Tables 7 and 8 show the results for GRF using binary data.

Each table shows the AUC (+ standard deviation) that is achieved for predicting someone will accept to buy a specific product online, given the media the person uses to search for supplier information online, offline or both; or given the media used for searching product information online, offline or both.

Tables 3 and 4 show interesting results. First, the AUC is systematically higher than 0,5, showing the data really has predictive power. With GRF, the AUC is typically even higher than 0,6 (when online media are used as input) and often even higher than 0,7. Comparing the AUC with the benchmark from [19], the results are promising. For example, the AUC that is achieved with the GRF and online supplier search information is systematically higher than the AUC reported in [19] (same holdout). Secondly, if we compare between online/offline channel usage information, having information on both Online and Offline channel usage is generally better than having information about online media usage only or offline media usage only. Furthermore, information about online channel preferences allows better predictions than offline channel preferences when using a 30% holdout. While largely irrelevant from a practical viewpoint, from a theoretical viewpoint it is interesting to note that there are still quite some cases where the AUC is larger than 60% when using information about offline channel preferences only. Thus, knowing the offline channel usage in early steps of the purchasing process can be helpful for predicting e-commerce adoption.

Zooming in on the GRF test results with 30% holdout, it is clear that having supplier search information is more predictive than product search information. Among the 7 product types, there are 5 cases where supplier search information provides better prediction than product search information with 1 case tied and 1 case the opposite. In the case of the Logistic regression, there are no systematically better results achieved when using supplier information than when using product information. Comparing the GRF results with the Logistic regression results, the GRF results are significantly better than those achieved with the Logistic regression.

If we compare the results within classification methods across holdout settings, increasing the holdout from 30% to 90% (table 3 versus 5, table 4 versus 6), it is obvious the predictive performance of logistic regression suffers significantly when the size of the training sample decreases. When only 10% of the data is available for training a model, the predictive results are far from satisfying as most of them are in the low 0,50s AUC. Meanwhile, GRF's performance decreases as well but the AUC is still considerably higher than that of the Logistic regression.

In contrast to what was noted for the GRF tests with a 30% holdout, we note that the supplier search information on average is no longer more valuable than the product search information when a 90% holdout is used with GRF.

Finally, we are also interested to know if we still get good (or even better) results if we simplify the ordinal data (1-5, most applicable to not applicable at all) into the binary choice (1 and 2 => relevant, 3-5 => not relevant) made by respondents for choice the channels. The results of GRF with a 30% holdout and a 90% holdout are shown in Table 7 and Table 8 respectively. The results are very promising as the results are generally even better than when using the rank of the media preferences. This is both the case, for the test with a 30% holdout and a 90% holdout. This result is important since in practice, companies rarely have a ranking of one's channel preferences. Our test results suggest that starting with very limited facts such as browsing history or other UGC (user generated contents) is sufficient to make relatively accurate predictions of e-commerce adoption.

Table 3 Results GRF with 30% holdout and ranked data

Product type	Prediction result reported in [19] as bench mark	Search for supplier information			Search for product information		
		Online	Offline	Online & Offline	Online	Offline	Online & Offline
1	0.62±0.05	0.65±0.05	0.55±0.05	**0.66±0.04**	0.64±0.05	0.50±0.05	0.64±0.05
2	0.58±0.04	0.68±0.04	0.53±0.04	**0.77±0.04**	0.65±0.04	0.59±0.05	0.74±0.04
3	0.63±0.05	0.64±0.05	0.57±0.06	**0.68±0.05**			
4	0.59±0.04	0.65±0.05	0.53±0.04	**0.71±0.04**	0.64±0.04	0.60±0.04	0.72±0.04
5	0.67±0.05	0.72±0.05	0.55±0.06	**0.75±0.04**			
6	0.63±0.04	0.73±0.04	0.51±0.05	**0.76±0.04**	0.66±0.04	0.55±0.04	0.71±0.04
7	0.58±0.05	0.62±0.06	0.56±0.05	**0.72±0.04**			
8	0.62±0.04	0.64±0.05	0.52±0.05	0.66±0.04	0.66±0.05	0.53±0.05	**0.71±0.04**
9	-	0.68±0.04	0.64±0.05	**0.75±0.04**			
10	0.66±0.08	0.69±0.08	0.65±0.07	**0.70±0.07**			
11	0.64±0.04	0.66±0.05	0.65±0.05	**0.69±0.04**	0.60±0.04	0.54±0.04	0.62±0.05
12	0.61±0.04	0.76±0.04	0.63±0.05	**0.80±0.03**	0.70±0.04	0.62±0.04	0.77±0.04
Average	0.62±0.05	0.68±0.05	0.57±0.05	**0.72±0.04**	0.65±0.04	0.56±0.04	0.70±0.04

Table 4 Results Logistic regression with 30% holdout and ranked data

Product type	Search for supplier information			Search for product information		
	Online	Offline	Online & Offline	Online	Offline	Online & Offline
1	0.58±0.06	0.54±0.05	0.61±0.05	0.64±0.05	0.48±0.05	0.62±0.05
2	0.65±0.04	0.54±0.04	0.75±0.04	0.63±0.04	0.61±0.04	0.72±0.04
3	0.54±0.07	0.53±0.05	0.60±0.06			
4	0.58±0.04	0.53±0.04	0.67±0.04	0.64±0.04	0.6±0.04	0.71±0.04
5	0.54±0.05	0.48±0.05	0.65±0.04			
6	0.63±0.04	0.48±0.05	0.72±0.04	0.58±0.04	0.53±0.04	0.67±0.04
7	0.61±0.06	0.57±0.04	0.71±0.04			
8	0.62±0.05	0.48±0.04	0.62±0.04	0.63±0.04	0.53±0.04	0.67±0.04
9	0.64±0.05	0.62±0.05	0.72±0.05			
10	0.61±0.09	0.61±0.09	0.62±0.07			
11	0.54±0.05	0.61±0.04	0.65±0.03	0.58±0.04	0.55±0.05	0.58±0.04
12	0.70±0.05	0.58±0.04	0.74±0.04	0.67±0.04	0.61±0.04	0.74±0.04
Average	0.60±0.05	0.55±0.05	0.67±0.05	0.62±0.04	0.56±0.04	0.67±0.04

Table 5 Result GRF with 90% holdout and ranked data

Product type	Search for supplier information			Search for product information		
	Online	Offline	Online & Offline	Online	Offline	Online & Offline
1	0.58±0.05	0.52±0.03	0.61±0.04	0.60±0.04	0.51±0.03	0.59±0.04
2	0.65±0.04	0.51±0.03	0.73±0.03	0.62±0.04	0.56±0.04	0.71±0.02
3	0.57±0.05	0.53±0.04	0.61±0.04			
4	0.61±0.05	0.51±0.03	0.67±0.03	0.62±0.04	0.56±0.04	0.68±0.04
5	0.67±0.05	0.53±0.03	0.66±0.05			
6	0.68±0.04	0.51±0.02	0.72±0.03	0.62±0.05	0.52±0.04	0.68±0.03
7	0.57±0.05	0.53±0.04	0.67±0.04			
8	0.61±0.04	0.50±0.02	0.61±0.04	0.62±0.05	0.51±0.03	0.65±0.04
9	0.66±0.04	0.59±0.05	0.70±0.04			
10	0.63±0.07	0.57±0.07	0.63±0.09			
11	0.61±0.05	0.62±0.04	0.63±0.04	0.58±0.05	0.51±0.04	0.58±0.04
12	0.73±0.04	0.59±0.04	0.77±0.02	0.68±0.03	0.58±0.04	0.74±0.03
Average	0.63±0.05	0.54±0.04	0.67±0.04	0.63±0.04	0.54±0.04	0.66±0.03

Table 6 Result Logistic regression with 90% holdout and ranked data

Product type	Search for supplier information			Search for product information		
	Online	Offline	Online & Offline	Online	Offline	Online & Offline
1	0.52±0.06	0.51±0.04	0.54±0.05	0.58±0.08	0.49±0.03	0.55±0.05
2	0.61±0.06	0.51±0.04	0.67±0.04	0.60±0.06	0.56±0.05	0.65±0.04
3	0.52±0.05	0.50±0.04	0.53±0.06			
4	0.53±0.05	0.51±0.03	0.59±0.05	0.60±0.05	0.55±0.05	0.64±0.05
5	0.52±0.06	0.50±0.03	0.56±0.05			
6	0.58±0.06	0.49±0.02	0.63±0.05	0.55±0.06	0.51±0.03	0.60±0.05
7	0.58±0.05	0.51±0.05	0.63±0.05			
8	0.58±0.06	0.49±0.03	0.56±0.05	0.56±0.07	0.50±0.04	0.60±0.05
9	0.61±0.05	0.57±0.05	0.60±0.07			
10	0.53±0.07	0.54±0.08	0.56±0.05			
11	0.51±0.05	0.56±0.04	0.58±0.05	0.54±0.06	0.53±0.04	0.55±0.05
12	0.65±	0.54±	0.66±0.05	0.63±	0.57±	0.66±0.05

	0.06	0.05		0.05	0.05	
Average	0.56±0.06	0.52±0.04	0.59±0.05	0.58±0.06	0.53±0.04	0.61±0.05

Table 7 Result GRF with 30% holdout (binary)

Product type	Search for supplier information			Search for product information		
	Online	Offline	Online & Offline	Online	Offline	Online & Offline
1	0.72±0.05	0.6±0.04	0.72±0.05	0.64±0.04	0.50±0.05	0.67±0.04
2	0.71±0.04	0.55±0.04	0.78±0.04	0.73±0.04	0.55±0.04	0.74±0.04
3	0.68±0.05	0.57±0.05	0.70±0.04			
4	0.67±0.05	0.55±0.04	0.72±0.04	0.68±0.04	0.57±0.04	0.72±0.04
5	0.73±0.04	0.63±0.05	0.77±0.04			
6	0.70±0.04	0.51±0.04	0.76±0.04	0.72±0.04	0.53±0.04	0.74±0.04
7	0.64±0.05	0.61±0.05	0.72±0.05			
8	0.71±0.05	0.58±0.04	0.76±0.04	0.73±0.04	0.52±0.04	0.74±0.03
9	0.75±0.04	0.64±0.05	0.79±0.04			
10	0.70±0.07	0.67±0.08	0.76±0.06			
11	0.71±0.04	0.61±0.04	0.72±0.04	0.64±0.04	0.55±0.04	0.66±0.04
12	0.74±0.04	0.65±0.04	0.82±0.03	0.78±0.04	0.63±0.04	0.79±0.04
Average	0.71±0.05	0.60±0.05	0.75±0.04	0.70±0.04	0.55±0.04	0.72±0.04

Table 8 Result GRF with 90% holdout (binary)

Product type	Search for supplier information			Search for product information		
	Online	Offline	Online & Offline	Online	Offline	Online & Offline
1	0.67±0.06	0.55±0.05	0.65±0.06	0.60±0.07	0.50±0.03	0.61±0.05
2	0.71±0.02	0.52±0.03	0.75±0.03	0.72±0.01	0.52±0.04	0.71±0.03
3	0.66±0.04	0.53±0.04	0.67±0.04			
4	0.67±0.03	0.53±0.03	0.69±0.03	0.66±0.04	0.53±0.04	0.69±0.03
5	0.73±0.04	0.60±0.04	0.74±0.02			
6	0.70±0.03	0.50±0.03	0.72±0.04	0.71±0.03	0.52±0.03	0.70±0.05
7	0.62±0.06	0.56±0.05	0.67±0.04			
8	0.71±0.05	0.54±0.05	0.73±0.03	0.72±0.03	0.5±0.03	0.71±0.04
9	0.76±0.02	0.58±0.06	0.76±0.04			
10	0.66±0.08	0.59±0.08	0.72±0.05			

11	0.70± 0.03	0.56± 0.06	0.69±0.03	0.63± 0.04	0.52± 0.04	0.62±0.04
12	0.73± 0.02	0.61± 0.04	0.80±0.02	0.77± 0.01	0.59± 0.05	0.77±0.02
Average	0.69± 0.04	0.56± 0.05	0.72±0.04	0.69± 0.04	0.53± 0.04	0.69±0.04

DISCUSSION

First, we contrast our results with those published in previous work [19]. The AUC that is achieved with the GRF function, using information about previous online behavior from the person, is higher than the AUC that was achieved in [19] using the social network information of people. Also, our approach does not need information on people's social ties. Rather information can be used on the person's browsing history. Just by investigating the use of Google and other search engines, online yellow pages and online auction sites such as e-bay, allows making good predictions of the e-commerce acceptance of a person. Such binary information can rather easily be gathered by companies (or bought from Internet ad serving firms such as DoubleClick). By making predictions about e-commerce acceptance, companies can make sure not to pay for publicity for online shops shown to people that are not open to buying that specific product online anyway.

Our research suggests that while there is a major emphasis nowadays on social influence (via online social networks for example) as a driver to increase e-commerce adoption (linked to the approach used in [19]), similarity between people based on personal channel preferences can be valuable in a different type of social network study as well, as shown here. Both, social influence and profile similarity, contribute to the observed correlation in networks. Neither of them should be ignored in e-commerce adoption research.

On first sight it seems surprising that using binary data generally gives better results than using ranked data. This fact could be caused by the fact that the ranking does not really matter so much in the head of the respondent. That is, whether some medium is deemed to be entirely inappropriate or only somewhat inappropriate to find a supplier does not really matter to the person, as the person will not use those channels anyway. In the end, some channels will be used and others will not be used, so that binary information very well reflects the relevant thoughts of the person. When creating rankings, people may be categorized as being dissimilar; while they actually show similar behavior in terms of the channel they are actually using (as measured by the binary variable).

One might say that it is unclear whether a person is looking for product information or a supplier when searching the Internet and that it thus could be hard to get the necessary data. Still, predictive performance is only slightly worse in case product search information is used with GRF with a 30% holdout. With binary data, no difference is noticed between using online supplier search information or online product search information so that it does not really matter (from an e-commerce acceptance prediction viewpoint) whether the person was actually searching for product information or a supplier.

Future research is needed to analyze the decision rules that are being generated during the training phase and that are used by the system when making predictions in the test phase. An analysis of those rules could give more insight into the reasoning and behavior of people.

The next step of our future research we will then investigate the difference between different products in more detail. Once we have analyzed the different rules, we can try to understand why the GRF function works better for one product and less good for another product. There is no simple explanation on first sight. For example, predictions for services are not systematically better or worse than predictions for products. While different rules are generated for different products, it is interesting from a theoretical viewpoint to find out why the GRF function performs differently for different products and to analyze whether the generated rules are very different for different products.

A limitation of our study is that we use a sample of students and their parents, so that we do not have a sample of the entire population. For example, the age category 25 to 40 is largely missed in our sample and future research could investigate the applicability of our approach in that part of the population. Still, we note it is appropriate to have used a student sample in our study, as students and their parents are really people who are consumers and who are really making decisions on buying offline or online. They are just not representative of the entire population but only of a (relevant) part of that population.

CONCLUSION

In this paper we use a semi-supervised learning model to predict e-commerce adoption of different products and services using information about the channels one deems appropriate to find product information or to find suppliers. Our results show that semi-supervised learning has advantages in terms of predictive performance compared to supervised learning, which demands a fully labeled dataset. The theoretical implication of our work is that both online and offline channel preferences are found to be predictive for e-commerce adoption. Using similarity in channel preferences between users, predictive models can utilize labeled and unlabeled data. The practical relevance of the work is that knowing users' preferred channel for information retrieval can be helpful to estimate the probability the person is interested to buy a specific product online. Besides studying real relations in a social network [19], connecting similar people in an artificial network is valuable to predict e-commerce acceptance.

REFERENCES

[1] Ajzen, I. 1985. "From Intentions to Actions: A Theory of Planned Behavior," in *Action Control,* J. Kuhl and J. Beckmann (eds.), Springer Berlin Heidelberg, pp. 11-39.

[2] Ajzen, I. 1987. "Attitudes, Traits, and Actions: Dispositional Prediction of Behavior in Personality and Social Psychology," in *Advances in Experimental Social Psychology,* B. Leonard (ed.), Academic Press, pp. 1-63.

[3] Ajzen, I. 1991. "The theory of planned behavior," *Organizational Behavior and Human Decision Processes* (50:2) 12, pp 179-211.

[4] Ajzen, I. 2002. "Perceived Behavioral Control, Self-Efficacy, Locus of Control, and the Theory of Planned Behavior1," *Journal of Applied Social Psychology* (32:4), pp 665-683.

[5] Bandura, A. 1986. *Social Foundations of Thought and Action: A Social Cognitive Theory*, (Prentice-Hall: Englewood Cliffs, NJ.

[6] Chen, C.-D., Fan, Y.-W., and Farn, C.-K. 2007. "Predicting electronic toll collection service adoption: An integration of the technology acceptance model and the theory of planned behavior," *Transportation Research Part C: Emerging Technologies* (15:5) 10, pp 300-311.

[7] Choudhury, V., and Karahanna, E. 2008. "The relative advantage of electronic channels: a multidimensional view," *MIS Q.* (32:1), pp 179-200.

[8] Demsar, J. 2006. "Statistical Comparisons of Classifiers over Multiple Data Sets," *J. Mach. Learn. Res.* (7), pp 1-30.

[9] Diaconis, P., and Graham, R. L. 1977. "Spearman's Footrule as a Measure of Disarray," *Journal of the Royal Statistical Society. Series B (Methodological)* (39:2), pp 262-268.

[10] Fawcett, T. 2006. "An introduction to ROC analysis," *Pattern Recogn. Lett.* (27:8), pp 861-874.

[11] George, J. F. 2004. "The theory of planned behavior and Internet purchasing," *Internet Research* (14 3), p 25.

[12] P. McCullagh, J. N. 1990. *Generalized Linear Models*, (Chapman & Hall: New York.

[13] Pavlou, P. A., and Fygenson, M. 2006. "Understanding and prediction electronic commerce adoption: An extension of the theory of planned behavior," *MIS Quarterly* (30:1), pp 115-143.

[14] Shmueli, G. 2010. "To Explain or to Predict?," *Statistical Science* (25:3), p 20.

[15] Shmueli, G., and Koppius, O. R. 2011. "Predictive analytics in information systems research," *MIS Q.* (35:3), pp 553-572.

[16] Socher, R., and Li, F.-F. Year. "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on2010, pp. 966-973.

[17] Tero Pikkarainen, K. P., Heikki Karjaluoto, Seppo Pahnila 2004. "Consumer acceptance of online banking: an extension of the technology acceptance model," Internet Research (14 3), pp 224 - 235.

[18] Verbraken, T., Goethals, F., Verbeke, W., and Baesens, B. 2012. "Using Social Network Classifiers for Predicting E-Commerce Adoption," in *E-Life: Web-Enabled Convergence of Commerce, Work, and Social Life,* M. Shaw, D. Zhang and W. Yue (eds.), Springer Berlin Heidelberg, pp. 9-21.

[19] Verbraken T., G. F., Verbeke W., Baesens B. 2013. "Predicting online channel acceptance using social network data," *Decision Support Systems*.

[20] Zhou, D., Huang, J., and Scholkopf, B. 2005. "Learning from labeled and unlabeled data on a directed graph," in *Proceedings of the 22nd international conference on Machine learning*, ACM: Bonn, Germany, pp. 1036-1043.

[21] Zhu, X., Ghahramani, Z., and Lafferty, J. Year. "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions," IN ICML2003, pp. 912-919.

[22] Zhu, X., and Goldberg, A. B. 2009. "Introduction to Semi-Supervised Learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning* (3:1) 2009/01/01, pp 1-130.