Learning User Purchase Intent from User-Centric Data

Rajan Lukose[1], Jiye Li[2], Jing Zhou[3], and Satyanarayana Raju Penmetsa[1]

[1] HP Labs, Palo Alto, California
{rajan_lukose,satyanarayana.raju}@hp.com
[2] Faculty of Computer Science and Engineering, York University
jiye@cse.yorku.ca
[3] Belk College of Business, UNC Charlotte
jzhou7@email.uncc.edu

Abstract. Most existing personalization systems rely on site-centric user data, in which the inputs available to the system are the user's behaviors on a specific site. We use a dataset supplied by a major audience measurement company that represents a complete user-centric view of clickstream behavior. Using the supplied product purchase metadata to set up a prediction problem, we learn models of the user's probability of purchase within a time window for multiple product categories by using features that represent the user's browsing and search behavior on all websites. As a baseline, we compare our results to the best such models that can be learned from site-centric data at a major search engine site. We demonstrate substantial improvements in accuracy with comparable and often better recall. A novel behaviorally (as opposed to syntactically) based search term suggestion algorithm is also proposed for feature selection of clickstream data. Finally, our models are not privacy invasive. If deployed client-side, our models amount to a dynamic "smart cookie" that is expressive of a user's individual intentions with a precise probabilistic interpretation.

1 Introduction

Clickstream data collected across all the different websites a user visits reflect the user's behavior, interests, and preferences more completely than data collected from one site. For example, one would expect that it would be possible to better model and predict the intentions of users who we knew not only searched for a certain keyword on a search engine S but also visited website X and the website Y, than if we knew only one of those pieces of information. The complete data set is termed user-centric data [8], which contains site-centric data as a subset. Most existing research on clickstream data analysis is based on site-centric data.

For the important task of personalization we seek to demonstrate rich, predictive user models induced from user-centric data, and quantify their advantages to site-centric approaches. We use a dataset supplied by a major audience measurement company that represents a complete user-centric view of clickstream

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 673–680, 2008.

behavior. The main contribution of our work is the first demonstration that accurate product category level purchase prediction modeling (regardless of the site of purchase) can be done from user-centric data. Using the supplied product purchase metadata to set up a prediction problem, we learn models of the user's probability of purchase within a time window for multiple product categories by using features that represent the user's behavior on all websites. Our model outperforms a reasonable and commercially meaningful baseline model learned from site-centric data restricted to a major search engine. We also propose a novel behaviorally (as opposed to syntactically) based search term suggestion algorithm which was an effective part of the feature selection strategy we used. Additionally, we explicitly consider the issue of prediction latency and show that even when predictions are made with long lead times, effective predictions can still be made. Finally, our models are not privacy invasive and we propose the idea of "smart cookies" motivated by our results. The success of our clickstream modeling approach should point the way to more personalization applications driven by clickstream modeling.

We first review the related background work in clickstream modeling and current research on personalization in Section 2. We then introduce our proposed online product purchase model and describe our experimental data in Section 3. Section 4 provides the experimental design and results.

2 Related Work

In the computer science literature, two main motivations have driven research on clickstream analysis: personalization and caching. Caching and prefetching to improve web server performance is obviously an important task and so site-centric clickstreams from web server logs have been analyzed to improve performance [4]. This line of work has emphasized the use of Markov models to predict page accesses. Despite a broad and deep interest, little direct work has been done on mining user-centric clickstream data for personalization. Site-centric personalization efforts have used clickstream analysis to cluster users [1,2] which enables site-specific content recommendation within user clusters. Additional work has been done in the marketing science literature [6] and [7]. User-centric clickstream data has been used in web personalization tasks such as personalized search [10], where clickstream data was part of the data used to help re-rank search results. Padmanabhan, *et al.* [8] demonstrated the predictive value of user-centric data versus site-centric data. Their work attempted to provide predictions of "purchase" or "no-purchase" at a given website (regardless of specific product category) based on user or site-centric data as inputs. In our work, we focus on the more widely useful and more difficult task of predicting specific *product category* purchases at *any* website. Furthermore, we consider search data as an important feature whose value as a prediction variable we are able to quantify and which was not used in this prior work.

3 Purchase Intent Model

This work focuses on developing general models that can effectively learn and predict users' online purchase intent. In these models, user-centric data is collected and stored in a database. After data preprocessing, features reflecting user online purchase intentions are constructed. The search terms that users input into general search engines, and the search terms they use on the leading online shopping stores are considered as indications of their purchasing interests (see [5] for more details). Then algorithms, such as decision trees, regression prediction algorithms are applied for predicting online purchase intent for various product categories on the processed data composed of the constructed features. We further explain the experimental dataset used, a search term suggestion algorithm, data preprocessing, feature construction and evaluations for the modeling process in the rest of this section.

3.1 Experimental Data

Nielsen Online MegaPanel data [1] is used as our testbed for purchase intent modeling. Nielsen is an online audience measurement company, which is a premier provider of high-quality internet data. The MegaPanel data is raw user-centric clickstream data, which includes, for example, online search behavior on leading search engines (such as Google, Yahoo) and shopping websites (such as Amazon, BestBuy). The data collection is processed to make the average customer's online behaviors consistent with a representative sampling of internet users. All personally identifying data is filtered from our dataset by Nielsen.

The data collected over 8 months amounted to approximately 1 terabyte from more than $100,000$ households. For each URL there are time stamps for each internet user's visit. Retailer transaction data (i.e. purchase metadata) contains more than 100 online leading shopping destinations and retailer sites. These data records show for a given user who makes a purchase online, the product name, the store name, the timestamp, the price and so on. Users' search terms can also be inferred from the URL data, which are collected from top search engines and comparison shopping sites (more details are given in [5]).

3.2 Behavior Based Search Term Suggestion Algorithm

Automatic discovery of relevant search terms can help construct features to distinguish buyers from non-buyers given a product category. The search terms users input into websites are indications of their purchasing intent, but it is a challenge to determine automatically which terms are relevant for a given product category. Current keyword suggestion tools are *syntactically* based, typically suggesting variations of queries that include a given seed search term. For example, for the purchase of "laptop", suggested keywords may include "laptops". Our approach is *behaviorally* based, does not use any information about syntactic

[1] http://www.nielsen-netratings.com/

Table 1. a) Top 10 Significant Terms for Sample Product Categories, b) Decision Table for Classifications

Apparel	Auto- motives	Books	Child BabyCare	Watch & Jewelry	Computer Hardware	Computer Software
granby	rotonda	amazon	thum	Seiko	dell	cafepress
coupon	civic	barnes	cravens	watches	dotnetnuke	panel
centreville	rotundra	books	aod	ebay	ati	hdtv
coupons	hfp	noble	mysterie811	movado	radeon	flat
shirts	ep3	goya	hohider	overstock.com	behringer	scripps
wrightsville	rechenberg	miquelon	strollers	watche	agp	plasma
clothing	bove	annapolis	pomade	xbox	laborer	kingman
pajamas	exhaust	diseases	dragonflies	Timex	hp	software
transat	switchers	autograph	toolady	Watchband	breakin	scroll
shirt	ifinder	griffie	gumball	Necklaces	blau	1080i

(a)

User ID	Condition Attributes 28 Features						Decision Attribute {buyer, non-buyer}
ID	G1a	G1b	. . .	G14c	G11	G16	{buyer, non-buyer}
1	Yes	2	. . .	7	5200	No	buyer
2	Yes	5	. . .	2	413	Yes	non-buyer
3	No	0	. . .	0	622	No	buyer
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
83,635	Yes	3	. . .	0	342	No	buyer

(b)

variation of queries, and does not even require seed terms. For example, related keywords under this method may include brand names such as “HP laptop”, “Dell” and “Lenovo” with no syntactic relationship to “laptop”.

We used the following algorithm to automatically generate a set of representative search terms. First, given a product category, we counted the frequencies of all the search terms observed from buyers over a certain period of time. Then we found which search terms are significantly different in frequency within the buyer population of our training data from the search terms which appear in the general population of buyers and non-buyers by using a *Z-value* test on each of the 26 product categories.

December 2005 data is used as our experimental data. We list the top 10 significant terms for sample product categories [2] in Table 1(a). This algorithm [5] was used for constructing useful features for our models in an automated way, but is also effective as a search term suggestion algorithm in more general contexts. For example, as can be seen in Table 1(a), this method identifies terms that do not include, and have no syntactic similarity to the word “watch” such as simple brand names like“seiko”, “movado”, and “timex” as well as misspellings such as “watche” and even other terms like “necklace”.

3.3 Feature Construction

We focus on constructing features that can reflect the users’ browsing and searching behaviors across multiple websites using user-centric data. There are 26 online product categories available in our experimental data. In this experiment, we consider the online purchasing product category to be personal computers, including both desktops and laptops.

We construct 28 features that are used in the following experiments for predicting purchase of personal computers. Such features include “whether searched laptop keywords before purchasing on Google”, “ # of sessions this user searched laptop keywords before purchasing”, “whether this user made a purchase (of any product category) in the past month” and so on (all features are listed in [5]). December 2005 data is used for this experiment.

[2] Note that random characters sequences are removed from the results.

4 Experiments

We discuss briefly the input data, experimental design, and evaluation metrics for the classification algorithms.

Input Data for Prediction. December 2005 data is used for this experiment. We consider the 28 features as condition attributes, and whether a person is a buyer or non-buyer for personal computers as the decision attribute. For a decision table $T = (C, D)$, $C = \{28 \text{ features}\}$, $D = \{\text{buyer, non-buyer}\}$. With $83,635$ users and 28 features, we create a decision table as shown in Table 1(b) as input to prediction algorithms for discovering users purchasing intent.

Experiment Design. For the complete data in the form of a decision table 83635×29 as shown in Table 1(b), we performed 10-fold cross validation through all the experiments.

Evaluation Metrics. We use the following evaluation metrics to evaluate classification performance. An individual can be classified as a buyer (denoted as P) or non-buyer (denoted as N). A classification is either correct (denoted as T) or false (denote as F). Thus, an individual who is an actual buyer but is classified as non-buyer is denoted by FN; an actual buyer and classified as a buyer is denoted as TP; an actual non-buyer but classified as buyer is denoted as FP; an actual non-buyer and classified as non-buyer is denoted as TN. Therefore, we have $Recall = \frac{TP}{TP+FN}$, $Precision = \frac{TP}{TP+FP}$, $TruePositiveRate = \frac{TP}{TP+FN}$, and $FalsePositiveRate = \frac{FP}{FP+TN}$.

4.1 Classification Experiments

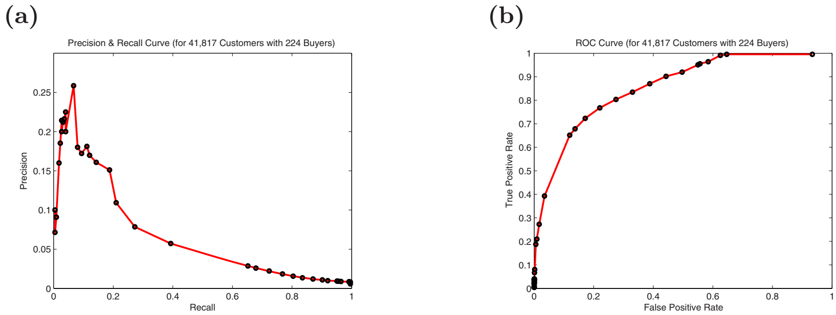
In order to accomplish the prediction task, we conducted the following experiments using classification algorithms including decision trees, logistic regression and Naïve Bayes.

Decision Tree. Decision trees can be used to construct classifiers for predictions. We assume only buyer or non-buyer as the two classes in our discussion. C4.5 decision tree [9] implementation is used for classification rule generation. We obtained precision 29.47%, and recall 8.37% for decision tree learning.

Logistic Regression. We use Weka's [3] logistic regression implementation for creating the classifier. By measuring the capabilities of each of the independent variables, we can estimate the probability of a buyer or non-buyer occurrence. The default cutoff threshold of predicting a buyer is p = 0.5. The precision is 18.52% and recall is 2.23%. Figure 1(a) shows the precision and recall curve for the user-centric classifier generated by logistic regression.

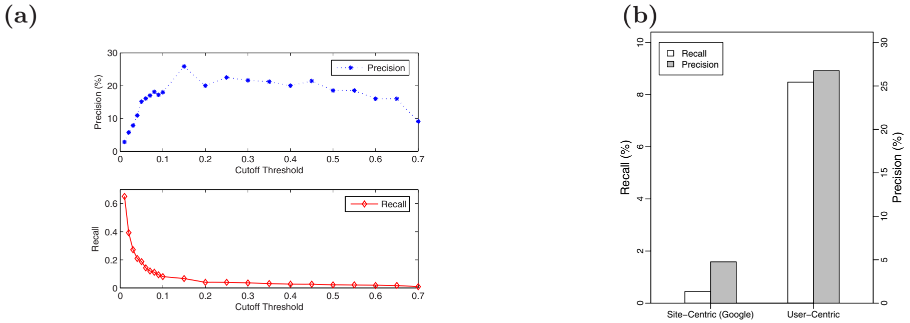
Figure 1(b) shows the ROC curve for the user-centric classifier generated by logistic regression. Figure 2(a) shows the tradeoff between the cutoff threshold and precision/recall for the user-centric classifier generated by logistic regression. This plot can be used for determining the suggested cutoff threshold in order to reach a satisfied precision and recall towards certain classification applications.

[3] Downloaded from http://www.cs.waikato.ac.nz/ml/weka/



Precision vs. Recall **ROC Curve**

Fig. 1. Experimental Results



Cutoff Threshold vs. Precision and Recall **Classifier Comparisons**

Fig. 2. Experimental Results

Naïve Bayes. Previous studies have shown that a simple Naïve Bayesian classifier has comparable classification performance with decision tree classifiers [3]. We use Weka's Naïve Bayes classifier implementation for our experiments [11]. We obtained the classification results as precision 3.52% and recall 23.2%.

Discussions. The classification experimental results demonstrate effective product level prediction. Classifiers can be created based on user-centric features to predict the potential buyers. From our experiment on predicting product purchases, we observed that decision tree algorithm can obtain the highest prediction precision. The branching nodes in the tree splitting a potential buyer and non-buyer can be detected and used for suggesting personalized relevant content. Logistic regression can be used as a flexible option to adjust the precision and recall for the classifiers.

4.2 Site and User-Centric Comparison Experiments

To help quantify the benefits of user-centric classifiers for this task, we compare the performance of a decision tree classifier based on 28 user-centric features to the best site-centric feature as a single classifier from a major search engine (i.e. "users who searched laptop keywords on Google before purchasing and searched more than one session"). The precisions for the user-centric and site-centric classifiers are 26.76% vs. 4.76%, and recall are 8.48% vs. 0.45%. The comparison figures is shown in Figure 2(b).

The result indicates that user-centric classifiers provide a much higher prediction precision (without loss of recall) than site-centric classifiers for predicting purchasing intent. Indeed, our discussions with industry experts indicate that even $\sim 5\%$ precision is an extremely good number in online marketing campaigns executed through search advertising. The fact that our models can increase precision, often with an increase in recall as well, demonstrates the rich value contained in user-centric data for widely applicable prediction problems.

4.3 Prediction Latencies

A key question for models of user intent is the prediction latency, defined as the period of time before the intended action that a prediction can be made. It may not be useful for many applications if good predictions can only be made over very short latent periods (e.g., a purchase prediction 10 seconds before it happens). To address this concern we performed latency experiments using November and December 2005 data. We used the feature "whether searched laptop keywords on all NNR before purchasing a personal computer", to make predictions using SQL aggregations. The experimental results indicate that 20.15% of computer transactions can be predicted by this feature. Among these predicted transactions, only 15.59% transactions have the latent period less than one day (we call this same-day-purchase) and 39.25% transactions have 1-7 days of latent period (we call this first-week-purchase). This experiment shows that online-shopping customers usually do not just come and immediately buy. They spend some time (mostly, more than one day) doing research before their final purchase decisions, which gives time to detect purchasing interests based on behaviors, make predictions, and suggest information.

4.4 Smart Cookies

Our results indicate that useful models of intent can be learned from offline panel data and could be deployed client-side through simple classification algorithms. Client-computed outputs such as "the probability that the user will purchase product type P within the next month" could be used as intentional signals for a variety of personalization tasks such as personalizing search or serving relevant advertising in a variety of contexts. These models need not be privacy invasive. A dynamic, intentionally expressive "smart cookie" could be one mechanism to deploy our models on the client-side. Whereas browser cookies often contain simple information such as identities, etc., we imagine an implementation using models such as the ones we have demonstrated which can augment the cookie data with *intentional* data. (See [5] for more details).

For example, Google now employs a feature called “web history”, which automatically collects and stores on central servers the entire clickstream of participating users. Presumably, some users would be more comfortable than others, and our methods show how to learn useful models from such data which can be deployed client-side on users who do not participate in such collection.

5 Conclusion

We demonstrated very effective product category level purchase prediction models (regardless of the site of purchase) for user-centric clickstream data. Comparison experiments show that the such models strongly outperform site-centric models, and predictions can be made ahead of time. Our models are fully automatable, and can be thought of as key enabling functionality for a “smart cookie” mechanism which could be deployed client-side and therefore would mitigate privacy concerns. It is worth noting that the baseline we established, the site-centric view of the search engine Google, was, by industry standards, quite good at predicting. Nevertheless, the user-centric models we created were able to outperform that important baseline by wide margins.

References

1. Banerjee, A., Ghosh, J.: Clickstream clustering using weighted longest common subsequences. In: Proc. of the Web Mining Workshop at the 1 st SIAM Conference on Data Mining, Chicago (2001)
2. Gunduz, S., Ozsu, M.: A web page prediction model based on click-stream tree representation of user behavior. In: KDD 2003, pp. 535–540 (2003)
3. Huang, J., Lu, J., Ling, C.X.: Comparing naive bayes, decision trees, and svm with auc and accuracy. In: ICDM 2003, p. 553 (2003)
4. Li, K., Qu, W., Shen, H., Wu, D., Nanya, T.: Two cache replacement algorithms based on association rules and markov models. In: SKG, p. 28 (2005)
5. Lukose, R., Li, J., Zhou, J., Penmetsa, S.R.: Learning user purchase intent from user-centric data. Technical report, Hewlett-Packard Labs (2008)
6. Moe, W.W., Fader, P.S.: Dynamic conversion behavior at e-commerce sites. Management Science 50(3), 326–335 (2004)
7. Montgomery, A.L., Li, S., Srinivasan, K., Liechty, J.C.: Modeling online browsing and path analysis using clickstream data. Marketing Science 23(4), 579–595 (2004)
8. Padmanabhan, B., Zheng, Z., Kimbrough, S.O.: Personalization from incomplete data: what you don’t know can hurt. In: KDD 2001, pp. 154–163 (2001)
9. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco (1993)
10. Teevan, J., Dumais, S.T., Horvitz, E.: Personalizing search via automated analysis of interests and activities. In: SIGIR 2005, pp. 449–456 (2005)
11. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann Publishers Inc., San Francisco (2005)