Noname manuscript No.
(will be inserted by the editor)

Forecasting User Visits for Online Display Advertising*

Suleyman Cetintas · Datong Chen · Luo Si

the date of receipt and acceptance should be inserted later

Abstract Online display advertising is a multi-billion dollar industry where advertisers promote their products to users by having publishers display their advertisements on popular Web pages. An important problem in online advertising is how to forecast the number of user visits for a Web page during a particular period of time. Prior research addressed the problem by using traditional time-series forecasting techniques on historical data of user visits; (e.g., via a single regression model built for forecasting based on historical data for all Web pages) and did not fully explore the fact that different types of Web pages and different time stamps have different patterns of user visits.

In this paper, we propose a series of probabilistic latent class models to automatically learn the underlying user visit patterns among multiple Web pages and multiple time stamps. The last (and the most effective) proposed model identifies latent groups/classes of i) Web pages and ii) time stamps with similar user visit patterns, and learns a specialized forecast model for each latent Web page and time stamp class. Compared with a single regression model as well as several other baselines, the proposed latent class model approach has the capability of differentiating the importance of different types of information across different

* This article significantly extends our previous efforts in (Cetintas et al, 2011a). Part of this work was done when the first author was visiting Yahoo! Labs.

Suleyman Cetintas
Dept. of Computer Sciences
Purdue University, West Lafayette, IN, 47907, USA
Tel.: +1 765 494 9165
Fax: +1 765 494 0739
E-mail: scetinta@cs.purdue.edu

Datong Chen
Yahoo! Labs
Santa Clara, CA, 95054, USA
E-mail: datong@yahoo-inc.com

Luo Si
Dept. of Computer Sciences and Statistics
Purdue University, West Lafayette, IN, 47907, USA
E-mail: lsi@cs.purdue.edu

classes of Web pages and time stamps, and therefore has much better modeling flexibility. An extensive set of experiments along with detailed analysis carried out on real-world data from Yahoo! demonstrates the advantage of the proposed latent class models in forecasting online user visits in online display advertising.

Keywords Forecasting · User Visits · Display Advertising

1 Introduction

Online advertising is one of the most profitable business models for Internet services. According to the Interactive Advertising Bureau (IAB), the total annual revenue of internet advertising in 2010 reaches a record level of $26 billion dollars, growing %15 from the previous year (IAB and PricewaterhouseCoopers, 2011). IAB further reports that "Consumers have shifted more of their time to digital media, watching television shows and movies online, and advertisers now accept this multifaceted medium as a key component for reaching their targets", and this is one of the main reasons for the steady increase in revenue across years. According to the same report, display related advertising is one of the major types of advertising, and had 38% of the whole advertising revenue in 2010 by itself.

In online graphical display advertising or online display advertising advertisers buy (explicitly or implicitly) targeted user visits from publishers in order to promote their products by displaying graphical (e.g., image, video) advertisements (ads) on popular Web pages. For instance, a sports shoes manufacturer may wish to purchase 50 million user visits by Males in Europe on Yahoo! Sports web sites during World Cup 2010. Online display advertising is related to, but different from sponsored search advertising (Aggarwal et al, 2006) or content match advertising (contextual advertising) (Broder et al, 2007). In sponsored search advertising advertisers bid for user-inputted keywords on search pages (Aggarwal et al, 2006; Richardson et al, 2007). In content match advertising or contextual advertising advertisers bid for clicks; and contextually relevant *text* ads (identified via text matching techniques) are shown to users (Broder et al, 2007). In both sponsored search or content match advertising, main goal of the advertisers is to obtain immediate clicks and purchases (conversions); and they typically pay per click (CPC or Cost Per Click) or conversion (CPA or Cost Per Action) (Agarwal et al, 2010a; Broder et al, 2007; Hatch et al, 2010; Yang et al, 2010). On the other hand; in display advertising, one of the main goals of the advertisers is to advertise their brands, and they typically buy impressions/ad views (CPM or Cost Per Mille - Cost Per Thousand Impressions) (Agarwal et al, 2010b; Alaei et al, 2009; Bharadwaj et al, 2010; Lahaie et al, 2008; Yang et al, 2010). CPM/Impression based pricing is the second most prevalent pricing model (ranking after performance based pricing), and had 33% of the whole advertising revenue in 2010 (IAB and PricewaterhouseCoopers, 2011).

An important problem in display advertising is how to forecast the count of user visits for a Web page (i.e., the number of impressions/ad views) during a particular period of time (e.g., day, hour, etc.). Over-forecasting user visit volumes may lead to undesired ad delivery outcomes such as missing an advertiser's goal and advertiser attrition. Similarly, under-forecasting the visit volumes may result in unsold user visits that often result in substantial revenue loss (Agarwal et al, 2010b; Feige et al, 2008).

Due to the increasing importance and popularity of computational advertising, there has been a large number of studies focusing on a variety of topics. One of the main topics studied extensively has been contextual ad matching, i.e., how to match ads to contextually relevant web pages (Agarwal et al, 2009; Broder et al, 2007; Chakrabarti et al, 2008; Hatch et al, 2010; Karimzadehgan et al, 2011; Lacerda et al, 2006; Murdock et al, 2007; Ribeiro-Neto et al, 2005). Another popular research direction has been how to estimate the click through rate (CTR) or impressions (i.e., forecasting) in sponsored search and content match (Agarwal et al, 2007; Richardson et al, 2007; Wang et al, 2009). Since the objectives and mechanisms of search related advertising and display advertising are very different, the techniques can not be applied directly to the forecasting problems in display advertising. Until recently, there has not been much attention paid to the forecasting system in display advertising. Prior research has mainly adopted traditional time-series forecasting techniques (Agarwal et al, 2010b; Cui et al, 2011; Shumway and Stoffer, 2007; Yang et al, 2010; Zellner and Tobias., 1999). Forecasting models are trained from historical user visits as a single regression model for all Web pages (assuming independence among different Web pages). In real online world, user visits among Web pages are not independent; and this results in groups of Web pages that have similar user visit patterns. For instance, one obvious factor shaping user visits is the physical structure of Web pages. A large proportion of users follow the navigation from a parent Web page, and visit the children Web pages step-by-step. Hence, some of the Web pages will have similar user visit patterns with each other, while having different user visit patterns than some other Web pages. Our prior work provided some preliminary research results to show that identifying (latent) groups of Web pages with similar user visiting behaviors, and specializing a different model for each group improves the forecast accuracy (Cetintas et al, 2011a). Similar to the Web pages, user visits are not independent of time (i.e., hour of day). Hence, some of the Web pages and time stamps will have similar user visit patterns with each other, while having different user visit patterns than some other Web pages and time stamps. To our best knowledge, prior approaches have not differentiated the groups of time stamps, and have not modeled groups of Web pages and time stamps with similar user visit patterns jointly.

In this paper, we learn user visit patterns hidden behind the online traffic of a large number of Web pages by using a series of novel probabilistic latent class models. In particular, we present three probabilistic latent class models that automatically identify different types/classes of Web pages and time stamps that share similar patterns of user visits from historical data, and build a regression model for each type/class of Web pages and time stamps for making accurate prediction. The first latent class model identifies groups of Web pages with similar user visit patterns, the second latent class model identifies groups of time stamps with similar user visit patterns, and finally the third model discovers groups of Web pages and time stamps that jointly share similar user visit patterns. The proposed models are compared to traditional time-series regression model as well as a large number of baselines that use past user visit information as the forecast. An extensive set of experiments along with detailed analysis on real-world proprietary data from Yahoo! shows the effectiveness of the proposed probabilistic latent class models as well as provides several insights about the performances of different models for forecasting user visit in online display advertising.

The rest of the paper is arranged as follows: Section 2 describes the proposed probabilistic latent class models. Section 3 discusses the experimental methodology, detailing the dataset, evaluation metric, and the baseline methods. Section 4 presents the experiment results along with detailed analysis; and finally Section 5 concludes this work.

2 Probabilistic Latent Class Models

This section first presents the probabilistic latent class model that automatically identifies different types/classes of Web pages and time stamps that jointly share similar patterns of user visits from historical data, and builds a regression model for each type/class of Web pages and time stamps for making accurate prediction. Then, it introduces a latent class model that only models latent Web page classes and a latent class model that only models latent time stamp classes as special cases of the joint latent class model.

In order to solve the forecasting problem in online display advertising, the prior work mainly adopted traditional time-series forecasting techniques that trained forecasting models from historical user visits as a single regression model for all Web pages and time stamps (Agarwal et al, 2010b; Cui et al, 2011; Shumway and Stoffer, 2007; Yang et al, 2010; Zellner and Tobias., 1999). Yet, different Web pages and time stamps have different user visiting behaviors. Therefore it is important to identify groups/classes of Web pages and time stamps with similar user visit behaviors, and specialize the forecasting model for each latent Web page and time stamp group/class. Although similar probabilistic models have been shown to be effective in many applications such as multimedia or web retrieval (Yan and Hauptmann, 2006), expert search (Fang et al, 2011), social network analysis (Cetintas et al, 2011b), etc.; prior forecasting approaches in online display advertising have (followed the traditional time-series forecasting techniques) not utilized the power of such probabilistic models. In this work, we propose a novel probabilistic latent Web page and time stamp class model that identifies groups/classes of Web pages and time stamps that share similar patterns of user visits, and learn a different forecasting model for each latent class.

Formally, let v_{st} be the user visit volume for a Web page/property s (Web pages or bucket of Web pages defined by a publisher) at time-stamp t, then the proposed probabilistic latent class model that identifies latent Web page classes as well as latent time stamp classes jointly can be described as follows:

$$P(s,t,v_{st}) = \sum_{z=1}^{N_z}\sum_{x=1}^{N_x} P(s)P(t)P(z|s)P(x|t)P(v_{st}|z,x) \tag{1}$$

where $P(s,t,v_{st})$ is the joint probability of s, t, and v_{st} (i.e., the probability of property s, time-stamp t of having user visit volume v_{st}), $P(s)$ and $P(t)$ are assumed to be uniform distributions, $P(z|s)$ denotes the conditional probability of a Web page latent class z given Web page s, $P(x|t)$ denotes the conditional probability of a time stamp latent class x given time stamp t, N_z is the number of latent Web page classes and N_x is the number of latent time stamp classes. The visit pattern in a class $P(v_{st}|z,x)$ can be modeled with a Laplace distribution as

follows:

$$P(v_{st}|z,x,f^{st},\lambda) = \frac{exp(-\frac{|v_{st}-\sum_i^K \lambda_{zxi} f_i^{st}|}{\beta})}{2\beta} \tag{2}$$

where f_i^{st} is the i^{th} feature for a Web page s and time stamp t pair (more information about the features can be found in Section 3.1), λ_{zxi} is the weight of latent Web page class z and latent time stamp class t for the i^{th} feature, and K is the number of features. It is important to note that it is possible to choose a different distribution to model the visit pattern in a class, such as the Gaussian distribution. Visits of users usually follow a similar pattern to Gaussian, yet have many more outliers. This work aims to minimize the absolute percentage error (that will be introduced as the evaluation metric in Section 3.3). That is, it is more important to forecast the majority of the user visit counts as accurately as possible, while leaving some room for some small number of (inevitable) outliers (i.e., outliers should not be allowed to change the model significantly). And due to the fact that Laplace distribution is more tolerant to (i.e., less effected by) outliers than the Gaussian distribution (Bishop, 2006; Boyd and Vandenberghe, 2004), we choose the Laplace distribution to model the user visit pattern in a class.

The parameters of the model in Eqn.(1) ($P(z|s)$, $P(x|t)$, λ) can be determined by maximizing the following data likelihood function:

$$\begin{aligned} & l\Big(P(z|s)\ ,\ P(x|t),\lambda\Big) \\ & \quad = \sum_{s,t} log\left(\sum_{z=1}^{N_z}\sum_{x=1}^{N_x} P(z|s)P(x|t)\frac{exp(-\frac{|v_{st}-\sum_i^K \lambda_{zxi} f_i^{st}|}{\beta})}{2\beta}\right) \end{aligned} \tag{3}$$

A typical approach to maximizing the data likelihood function above is to use the Expectation-Maximization (EM) algorithm (Dempster et al, 1977), which can obtain a local optimum of log-likelihood by iterating E-step and M-step until convergence. The E-step can be derived as follows by computing the posterior probability of z and x.

$$P(z,x|s,t,v_{st}) = P(z|s)P(x|t)\frac{exp(-\frac{|v_{st}-\sum_i^K \lambda_{zxi} f_i^{st}|}{\beta})}{2\beta} \tag{4}$$

By optimizing the auxiliary Q-function, we can derive the following M-step update rules:

$$P(z|s)^* \propto \sum_{x,t} P(z,x|s,t,v_{st}) \tag{5}$$

$$P(x|t)^* \propto \sum_{z,s} P(z,x|s,t,v_{st}) \tag{6}$$

$$\lambda^*_{zx\cdot} \propto \arg\max_{\lambda_{zx\cdot}} \sum_{s,t} P(z|s)P(x|t)\Big(-|v_{st}-\sum_i^K \lambda_{zxi} f_i^{st}|\Big) \tag{7}$$

Details about the derivation of E & M-step update rules in general can be found in (Bishop, 2006; Dempster et al, 1977). Eqn.(7) is differentiable, and can be solved with gradient descent solvers. In particular, we use the Quasi-Newton method.

This joint latent class model will be referred as *Latent_ST_Mod* for the rest of the paper.

In order to better understand the performance of the proposed probabilistic latent Web page and time stamp class model, three sub-models are also constructed: a probabilistic latent Web page class model, a probabilistic latent time stamp class model, and a simple regression (or traditional time-series forecasting) model. Note that when *Latent_ST_Mod* uses only one latent time stamp class and more than one latent Web page classes (i.e., $N_t = 1$ and $N_z > 1$), the latent Web page and time stamp class model degenerates to the latent Web page class model that only models latent Web page classes; and this latent class model will be referred as *Latent_S_Mod*. Similarly when *Latent_ST_Mod* uses only one latent Web page class and more than one latent time stamp classes (i.e., $N_z = 1$ and $N_t > 1$), the model degenerates to the latent time stamp class model that only models latent time stamp classes; and this latent class model will be referred as *Latent_T_Mod*. An extreme case is when *Latent_ST_Mod* uses only one latent Web page class and only one latent time stamp class (i.e., $N_z = 1$ and $N_t = 1$). In this case only the Laplace regression power is employed. We particularly report this case as *Laplace_Regr* in the experiments as one of the baselines. Note that the *Latent_ST_Mod*, *Latent_S_Mod*, *Latent_T_Mod* models and the *Laplace_Regr* model are run on the log-scaled (base-e) count data, and the estimated count is rescaled back for comparison with the raw user visit counts during evaluation. It should also be noted that, although the time that is required to train the models differs depending on model complexity, this is not of any practical importance as the training is done off-line and as all models can make their estimations (i.e., $v_{st}^{forecast} = \sum_{z,x} P(z|s)P(x|t)(\sum_i^K \lambda_{zxi} f_i^{st})$) instantly once the parameters of the models (i.e., $P(z|s), P(x|t), \lambda_{zxi}$) are learned.

3 Experimental Methodology

3.1 Dataset

Experiments are conducted on 1 month user visit logs of tens of millions of users on thousands of Yahoo! properties (i.e., Y! Web pages or bucket of Web pages such as Yahoo! Sports, Yahoo! Finance, etc. defined by a publisher). Features are extracted from historical counts of user visits of each property from its past week history, and user visits in the same hour are aggregated together (i.e., the forecast granularity is a time period of one hour). Data from the second and third weeks are used for training the models, and data from the last week are used for testing. Specifically, starting from the first hour of the second week, we extract the first 4 features as the average of visit volumes of the same hour-of-the-day in the past $1, 3, 5, 7$ days. In other words, we extract the historical count of user visits for a particular property for a specific hour of a specific day from the same hour in a window of 7 days in the history. For instance, the features for the hour, 9:00pm-9:59pm on Jan.10^{th} of a particular property s, are extracted from the visits on s during 9pm-9:59pm on days between Jan.9^{th}, 8^{th},..., Jan.3^{rd}. Note that the first week data are only used while extracting this first set of features for the second week. We extract the second 4 features as the average number of visits in the past $1, 3, 6, 9$ hours. Finally, a binary feature indicating whether the day is a weekend day is also used to

capture weekend holiday seasonality. In online display advertising, the main focus is on the most visited properties (i.e., Web pages or bucket of Web pages) that generate the main impressions/ad views. Therefore the number of user visits for the most visited i) 500, and ii) 1000 properties are computed (for each hour), and the corresponding datasets are referred as Top500Prop and Top1000Prop respectively. It is important to note that, the set of 500/1000 properties are, as noted before, not individual Web pages, but Web pages or buckets of Web pages such as Y! Sports, Y! Mail, etc. defined by a publisher. Therefore the set of the most visited 500 properties (i.e., Top500Prop) covers more than the majority of the important properties, and the set of 1000 most visited properties is definitely a safe set of properties that does not leave behind any property which is interesting in terms of display advertising. The Top500Prop and the Top1000Prop datasets have around $156K$ and $306K$ training data instances; and around $83K$ and $164K$ test data instances respectively.

It is important to note that historical count information is the standard source of information for forecasting problems in general, and has been used as the information source for the forecasting task in display advertising by prior works (Agarwal et al, 2010a; Cetintas et al, 2011a; Yang et al, 2010). Although this work utilizes features that are extracted from the historical counts of user visits of properties, other sources of information have also been considered. One such source is the Yahoo! Ontology that provides a hierarchical index of categories of Web pages based on page content (Labrou and Finin, 1999). Since the Yahoo! Ontology provides which properties (buckets of Web pages) are similar in their content (i.e., are of the same category/topic) and has parent-child relationship, both of which have high potential to affect the user visit patterns; it has high potential to be useful for the forecasting task. It should be noted that we experimented integrating the Y! Ontology information (with a modified model) especially to better identify latent groups of properties that share similar user visit patterns. Yet, we did not observe any improvements in model accuracy (and therefore these results as well as the modified model are not reported). This can be explained by the fact that the ontology information, which is constructed using the long-term page content information; provides too coarse signals to be useful, and is shadowed by the dynamically updated, fine-grained historical count information.

3.2 Baselines

The proposed probabilistic latent class model is compared to 3 types of baselines. The first baseline follows a simple forecasting approach, and uses the average of past visit volume as the forecast of the coming hour. We use the 8 features as a set of 8 Web-page-independent baseline forecasts $B\_LastNDay$ for $N \in (1, 3, 5, 7)$ and $B\_LastNHour$ for $N \in (1, 3, 6, 9)$.

The second baseline of this work, namely $BB\_PropSpec$, is similar to the first one, but allows each web page to have its own best model selected from the 8 features. The best feature is selected in the training data, and tested on the testing data.

The third baseline is a traditional time-series regression model. We use the existing 8 historical count features along with the binary flag feature that indi-

cates weekends, and perform regression with Laplace distributions. This model is referred as *Laplace_Regr*.

3.3 Evaluation Metric: Absolute Percentage Error (APE)

The forecasting error is measured by the absolute percentage error (APE) between forecast and truth:

$$APE = \frac{|v_{st}^{forecast} - v_{st}|}{v_{st}} \tag{8}$$

where $v_{st}^{forecast}$ is the forecast and v_{st} is the actual visit count. In Tables 1-4, we reported the mean of the absolute percentage error of the corresponding models for the Top500Prop and Top1000Prop datasets. Note that absolute percentage error (or relative error) is a standard metric used for evaluating forecasting performance in display advertising, and has been used as the evaluation metric in recent related works (Agarwal et al, 2010a; Cetintas et al, 2011a; Cui et al, 2011). Note that APE, unlike traditional metrics such as MAE or RMSE, normalizes the estimation error with the actual visit count, and therefore calculates an error depending on the visit volume of a property, which is very important for the evaluation of forecasting accuracy in online display advertising. For instance, an estimation error of 1000 (i.e., $|v_{st} - v_{st}^{forecast}|$) should not be punished too much in case of a highly visited property (e.g., $v_{st} = 100000$ and $v_{st}^{forecast} = 99000$) whereas it should be significantly punished for a property with much less visit counts (e.g., $v_{st} = 1000$ and $v_{st}^{forecast} = 2000$). For this scenario, the APE is 0.01 and 1 respectively for the two cases, whereas the absolute error is 1000 for both; which clearly shows the importance of the normalization factor.

In order to protect the confidential information from the company, the actual errors are normalized with (i.e., divided by) the error of $B\_Last1Hour$ baseline on the Top1000Prop dataset (note that the actual error is in the range 0.5-1), and only the normalized errors are reported for relative comparison.

4 Experiment Results

This section presents the experimental results of the probabilistic latent class models that are proposed in Section 2 in comparison to the various baselines that are presented in Section 3.2. All the models were evaluated on the dataset described in Section 3.1. An extensive set of experiments are conducted to address the following sets of questions:

1. How effective are the baseline methods (i.e., B_Last{1,3,5,7}Days, B_Last{1,3,6,9}Hours, BB_PropSpec, and LaplaceRegr models) in comparison to each other? (Section 4.1)
2. How effective is the probabilistic latent Web page class model (i.e., Latent_S_Mod) in comparison to several baselines? How sensitive is the latent class model with respect to the number of latent Web page classes? (Section 4.2)
3. How effective is the probabilistic latent time-stamp class model (i.e., Latent_T_Mod) in comparison to several baselines? How sensitive is the latent class model with respect to the number of latent time stamp classes? (Section 4.3)

Table 1 (Normalized) Results of several baselines in comparison to each other. The † and ‡ symbols indicate statistical significance with p-value < 0.001 with each model in comparison to $B\_Last1Day$ and $B\_Last1Hour$ respectively. The performance is evaluated by the mean of the Absolute Percentage Error (APE). This set of results can also be found in our previous work in (Cetintas et al, 2011a).

Methods	Top500Prop	Top1000Prop
B_Last7Days	2.293	2.385
B_Last5Days	2.288	2.451
B_Last3Days	2.091	2.318
B_Last1Day	1.195	1.794
B_Last9Hours	3.007	3.493
B_Last6Hours	2.116	2.577
B_Last3Hours	$1.077^{\dagger}$	$1.493^{\dagger}$
B_Last1Hour	$0.614^{\dagger}$	$1^{\dagger}$
BB_PropSpec	$0.987^{\dagger}$	$1.229^{\dagger}$
Laplace_Regr	$0.545^{\dagger,\ddagger}$	$0.893^{\dagger,\ddagger}$

4. How effective is the probabilistic latent Web page and time-stamp class model (i.e., Latent_ST_Mod) in comparison to several baselines as well as the latent Web page class model (i.e., Latent_S_Mod) and the latent time stamp class model (i.e., Latent_T_Mod)? How sensitive is the latent class model with respect to the number of latent Web page and time stamp classes? (Section 4.4)
5. How robust are the baseline methods B_Last1Day, B_Last1Hour, Laplace_Regr and the probabilistic latent Web page, latent time-stamp, and latent Web page and time stamp class models (i.e., Latent_S_Mod, Latent_T_Mod, and Latent_ST_Mod) in forecasting the user visits of all properties? What are the strenghts and weaknesses of each method? (Section 4.5)

4.1 The Performance of Several Baselines

The first set of experiments in this section was conducted to evaluate the performances of the various baseline methods in comparison to each other (i.e., B_Last{1,3,5,7}Days, B_Last{1,3,6,9}Hours, BB_PropSpec, and LaplaceRegr models). The details about these approaches are given in Section 3.2.

It can be seen from Table 1 that among the four baselines that follow the simple forecasting approach of using the average of past visit volume of the previous days as the forecast for the coming hour, i.e. the B_Last{1,3,5,7}Days, $B\_Last1Day$ is by far the best performing baseline, followed by $B\_Last3Day$. It can also been seen that the baselines that use more-up-to-date data mostly perform better, which shows that more up-to-date average visit volume information is a better estimate of the current visit volume.

Among the baselines that use the average of past visit volume of the previous hours as the forecast for the coming hour, i.e. the B_Last{1,3,6,9}Hours, $B\_Last1Hour$ is the best performing baseline, followed by $B\_Last3Hour$,

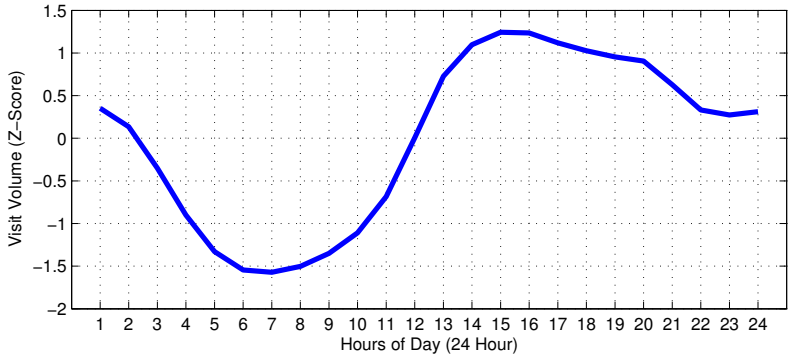

Fig. 1 User visit traffic across different hours of the day. Standard scores (z-scores) of the traffic accumulated for all properties are reported.

*B_Last*6*Hour*, and lastly *B_Last*9*Hour*. Similar to the observation above, the most up-to-date information the baseline uses the higher the accuracy becomes. Note that *B_Last*1*Hour* is also the best performing model out of all 8 baselines that use the average volume from the past as the forecast for the coming hour (i.e., B_Last{1,3,5,7}Days and B_Last{1,3,6,9}Hours), and achieves statistically very significant (paired t-tests) improvements (with p-value much less than 0.001) over all other methods. This shows that the user visits in the last hour is the most relevant to the current hour, which is totally consistent with the common sense. An important observation is that although *B_Last*6*Hours* and *B_Last*9*Hours* use more up-to-date information than the B_Last{1,3,5,7}Days baselines, they are outperformed by *B_Last*1*Day* and *B_Last*3*Days*. This is due to the fact that there is a significant change in the counts of user visits across different hours throughout the day for most properties. Figure 1 plots the distribution of normalized hourly traffic (accumulated for all properties) for hours of day. It can be seen that the traffic starts decreasing after midnight and reaches peak low around 6-7AM whereas it starts to increase with the morning hours and reaches its peak high at 3PM. It is clear that using the user visit information from the last few hours is highly likely to be a good forecast for the next hour traffic justifying the performance of *B_Last*1*Hour*. On the other side, using a longer history such as the last 6 or 9 hours average may not be a good way of estimating the coming hour depending on the hour to be forecasted. For instance, using last 6 or 9 hours of data will highly overestimate the traffic for the peak low hours of 6 or 7AM, and will highly underestimate the traffic for the peak high hours of 3PM, etc. Yet, it should be noted that Figure 1 shows the traffic accumulated for all properties, and for individual properties the traffic trend may be different than the global trend.

Table 1 also shows that *BB_PropSpec* performs significantly better (with p-value much less than 0.001) than all approaches that use the average of past visit volume as the forecast for the coming hour except *B_Last*1*Hour*. Selecting the best model for each Web page overfits the training data, and generates more forecasting errors even in comparison to the simple baseline *B_Last*1*Hour*. Potential improvements can be achieved by following a direction between these two types of baseline approaches.

Table 2 (Normalized) Results of the proposed probabilistic latent Web page class model (i.e., Latent_S_Z_Mod where Z is the number of latent Web page classes) in comparison to the best performing baselines. The †, ‡, § symbols indicate statistical significance with p-value < 0.001 with each model in comparison to $B\_Last1Day$, $B\_Last1Hour$, and $Laplace\_Regr$ respectively. The performance is evaluated by the mean of the Absolute Percentage Error (APE). A small part of these results can also be found in our previous work in (Cetintas et al, 2011a).

Methods	Top500Prop	Top1000Prop
B_Last1Day	1.195	1.794
B_Last1Hour	$0.614^{\dagger}$	$1^{\dagger}$
BB_PropSpec	$0.987^{\dagger}$	$1.229^{\dagger}$
Laplace_Regr	$0.545^{\dagger,\ddagger}$	$0.893^{\dagger,\ddagger}$
Latent_S_5_Mod	$0.483^{\dagger,\ddagger,\S}$	$0.828^{\dagger,\ddagger,\S}$
Latent_S_{10}_Mod	$0.484^{\dagger,\ddagger,\S}$	$0.828^{\dagger,\ddagger,\S}$
Latent_S_{15}_Mod	$0.482^{\dagger,\ddagger,\S}$	$0.828^{\dagger,\ddagger,\S}$
Latent_S_{20}_Mod	$0.482^{\dagger,\ddagger,\S}$	$0.826^{\dagger,\ddagger,\S}$
Latent_S_{25}_Mod	$0.481^{\dagger,\ddagger,\S}$	$0.826^{\dagger,\ddagger,\S}$
Latent_S_{30}_Mod	$0.486^{\dagger,\ddagger,\S}$	$0.828^{\dagger,\ddagger,\S}$
Latent_S_{40}_Mod	$0.481^{\dagger,\ddagger,\S}$	$0.826^{\dagger,\ddagger,\S}$

Finally, it can be seen in Table 1 that the *Laplace_Regr* significantly outperforms (with p-value much less than 0.001) all previously compared approaches. This clearly shows that combining the different information from past user visits intelligently along with the binary flag that indicates weekends, is more effective than using only a specific type of historical count information. This can be explained by the fact that different types of historical count information provide different signals for the user visits for a specific hour, and utilizing all of these signals helps *Laplace_Regr* achieve better forecast accuracy.

4.2 The Performance of the Probabilistic Latent Web Page Class Model (Latent_S_Mod)

The second set of experiments was conducted to i) evaluate the performance of the proposed probabilistic latent Web page class model (i.e., Latent_S_Mod) in comparison to all baselines, and ii) to check the robustness of the latent class model with respect to different number of latent Web page classes. The details about the latent Web page class model is given in Section 2, and the details about the baselines are given in Section 3.2.

Table 2 shows that the proposed latent Web page class model (i.e., *Latent_S_Mod*) significantly outperforms (with p-value much less than 0.001) all the baseline approaches by modeling the latent Web page classes that provide much higher modeling flexibility leading to its superior performance. This explicitly shows that differentiating the Web pages with different user visit patterns, and specializing the forecast model for different types/classes of Web pages that share similar patterns of user visits is important for achieving higher forecast accuracy.

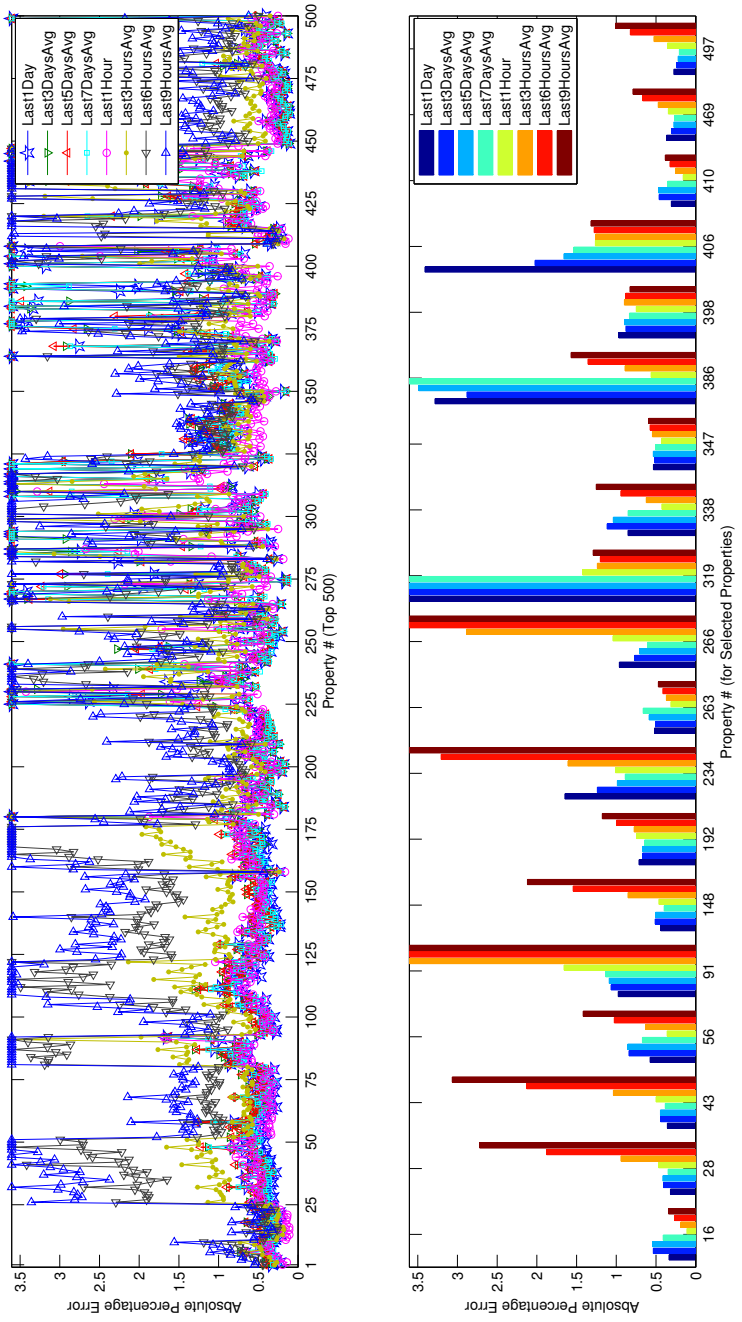

Fig. 2 Predictiveness of different types of past visit volume information for the top 500 most visited properties in the upper subfigure, and for a sampled set of properties (in more detail) in the lower subfigure. Different types of past visit volume information have different predictive value for different properties resulting in groups of properties with similar visit volume patterns.

Table 3 (Normalized) Results of the proposed probabilistic latent time stamp class model (i.e., Latent_T_X_Mod where X is the number of latent time stamp classes) in comparison to the best performing baselines.The †, ‡, § symbols indicate statistical significance with p-value < 0.001 with each model in comparison to $B\_Last1Day$, $B\_Last1Hour$, and $Laplace\_Regr$ respectively. (Note that § is never achieved). The performance is evaluated by the mean of the Absolute Percentage Error (APE).

Methods	Top500Prop	Top1000Prop
B_Last1Day	1.195	1.794
B_Last1Hour	$0.614^{\dagger}$	$1^{\dagger}$
BB_PropSpec	$0.987^{\dagger}$	$1.229^{\dagger}$
Laplace_Regr	$0.545^{\dagger,\ddagger}$	$0.893^{\dagger,\ddagger}$
Latent_T_3_Mod	$0.552^{\dagger,\ddagger}$	$0.906^{\dagger,\ddagger}$
Latent_T_4_Mod	$0.545^{\dagger,\ddagger}$	$0.904^{\dagger,\ddagger}$
Latent_T_5_Mod	$0.554^{\dagger,\ddagger}$	$0.906^{\dagger,\ddagger}$
Latent_T_6_Mod	$0.554^{\dagger,\ddagger}$	$0.908^{\dagger,\ddagger}$
Latent_T_{10}_Mod	$0.558^{\dagger,\ddagger}$	$0.903^{\dagger,\ddagger}$
Latent_T_{15}_Mod	$0.555^{\dagger,\ddagger}$	$0.901^{\dagger,\ddagger}$

It can also be seen from Table 2 that the performance of the Latent_S_Mod slightly changes with respect to the number of latent Web page classes. It is interesting to note that the model achieves the best results for the Top500Prop dataset around 15, 20 latent classes while it achieves the best results for the Top1000Prop dataset around 20, 25 latent classes. This is due to the fact that a dataset with more properties is more likely to have higher number of property groups with similar user visit patterns, more latent Web page groups help the model better capture these patterns.

Figure 2 plots the predictiveness of different types of past user visit volume information to the most visited 500 properties (i.e., Top500Prop dataset) in the upper subfigure, and for a sampled set of properties (in more detail) in the lower subfigure. It can be seen that there are several groups of Web pages for which different types of past visit information are more predictive than others. For instance, past visit information from the previous hour can be observed as one of the most important past visit information across all properties, and specifically the most informative for the properties between ranks 1 and 25 (e.g., property 16), ranks right after 50 (e.g., property 56), ranks between 325 and 350 (e.g., property 338). The past visit information from the previous day (i.e., Last1Day) is the most predictive for the properties with ranks between 25 and 50 (e.g., property 28), and between the ranks 75 and 175 (e.g., property 91), competing with other information as well. On the other hand, the same type of information is one of the worst predictor for the properties between ranks 375 and 400 (e.g., property 406). Similarly the past visit information from the past 7 days (i.e., Last7DaysAvg) is a very important predictor for many properties, and is the most predictive information for properties between ranks 180 and 220 (e.g., property 192), between 230 and 275 (e.g., property 266), and after 450 (e.g., properties 469 and 497). Yet, the same type of information is one of the worst predictive types of information for properties between ranks 375 and 400 (e.g., property 386) as well as some others. Similarly,

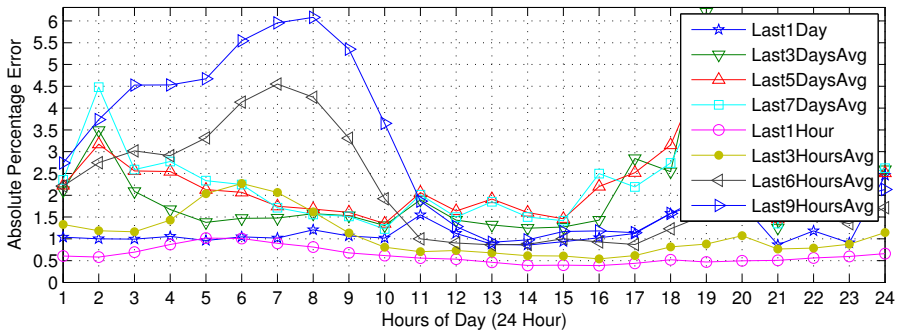

Fig. 3 Predictiveness of different types of past visit volume information across hours of the day. The types of past visit volume information that are more predictive for different hours of the day follow a similar pattern across all hours making it hard to identify different time stamp groups with similar visit volume patterns.

past visit information from the past $1, 3, 5, 7$ days is overall a much better predictor than the past visit information from the past $1, 3, 6, 9$ hours for many properties such as 28, 43, 91, 148, 266, 497, is a much worse predictor for many properties such as 319, 386, 406, and has comparable performance for others such as 16, 263, 347, 410. Therefore it is clear that the importance (i.e., predictiveness) of different types of past user visit information are different for different properties. A global model learned for all properties (e.g., Laplace_Regr) will weight the importance of all types of past user visit information in the global scale, and therefore will be conservative about the types of information such as Last1Day or Last7DaysAvg whose predictiveness fluctuate significantly across different properties. On the other side, the probabilistic latent Web page class model (i.e. Latent_S_Mod) has the capability of differentiating the groups of Web pages for which different types of past user visit information are more important than other types of information, and learns a specialized model for each of the Web page groups (i.e., latent classes). Therefore, it significantly outperforms the Laplace_Regr model as well as the other baselines via its higher modeling flexibility.

4.3 The Performance of the Probabilistic Latent Time Stamp Class Model (Latent_T_Mod)

The third set of experiments was conducted to i) evaluate the performance of the proposed probabilistic latent time stamp class model (i.e., Latent_T_Mod) in comparison to all baselines, and ii) to check the robustness of the latent class model with respect to different number of latent time stamp classes. The details about the latent time stamp class model is given in Section 2, and the details about the baselines are given in Section 3.2.

Table 3 shows that the proposed probabilistic latent time stamp class model (i.e., *Latent_T_Mod*) performs comparably with the Laplace_Regr approach, and is not able to achieve performance improvements despite the higher modeling flexibility it has though its latent time stamp classes. This is due to the fact that the

Table 4 (Normalized) Results of the proposed probabilistic latent Web page and time stamp classes model (i.e., Latent_$S_Z T_X$_Mod) in comparison to the best performing baselines as well as the Latent_S_Z_Mod and the Latent_T_X_Mod, where Z is the number of latent Web page classes and T is the number of latent time stamp classes. The †, ‡, §, ¶ symbols indicate statistical significance with p-value < 0.001 with each model in comparison to $B\_Last1Hour$, $Laplace\_Regr$, $Latent\_T_3\_Mod$ and $Latent\_S_{20}\_Mod$ respectively. The performance is evaluated by the mean of the Absolute Percentage Error (APE).

Methods	Top500Prop	Top1000Prop
B_Last1Day	1.195	1.794
B_Last1Hour	0.614	1
BB_PropSpec	0.987	1.229
Laplace_Regr	$0.545^{\dagger}$	$0.893^{\dagger}$
Latent_S_{20}_Mod	$0.482^{\dagger,\ddagger,\S}$	$0.826^{\dagger,\ddagger,\S}$
Latent_T_3_Mod	$0.552^{\dagger}$	$0.906^{\dagger}$
Latent_S_5T_3_Mod	$0.464^{\dagger,\ddagger,\S,\P}$	$0.812^{\dagger,\ddagger,\S,\P}$
Latent_$S_{10}T_3$_Mod	$0.461^{\dagger,\ddagger,\S,\P}$	$0.811^{\dagger,\ddagger,\S,\P}$
Latent_$S_{15}T_3$_Mod	$0.453^{\dagger,\ddagger,\S,\P}$	$0.804^{\dagger,\ddagger,\S,\P}$
Latent_$S_{20}T_3$_Mod	$0.463^{\dagger,\ddagger,\S,\P}$	$0.798^{\dagger,\ddagger,\S,\P}$
Latent_$S_{25}T_3$_Mod	$0.456^{\dagger,\ddagger,\S,\P}$	$0.805^{\dagger,\ddagger,\S,\P}$

predictiveness of the past user visit information across different time stamps is not significantly different than each other. Figure 3 plots the predictiveness of different types of past user visit volume information across hours of day (for the Top500Prop dataset). It can be seen that for all hours of day the past user information from the previous hour (i.e., Last1Hour) is the most important information to forecast the traffic for the coming hour. Similarly, the past user visit information from the previous 3 hours (i.e., Last3HoursAvg) is the second most important type of past user visit information between hours 9 and 24, and is the third most important type of information between hours 1 and 4. Similarly, the past user visit information from the previous day (i.e., Last1Day) is the second most important type of information between hours 1 and 9 with some small fluctuations afterwards. Overall, these three types of past user visit information is roughly the three most important types of information, and this trend does not significantly change across different hours, unlike the change these three information types had across different Web pages (as mentioned in the previous section). Therefore, specializing the learning across different time stamps does not help the model improve the forecast accuracy; and $Latent\_T\_Mod$ is not able have any improvements over the baseline LaplaceRegr modeling as well as the other baselines.

4.4 The Performance of the Probabilistic Latent Web Page and Time Stamp Class Model (Latent_ST_Mod)

The last set of experiments was conducted to evaluate the performance of the proposed probabilistic latent Web page and time stamp class model (i.e., Latent_ST_Mod) in comparison to all baselines as well as the latent Web page

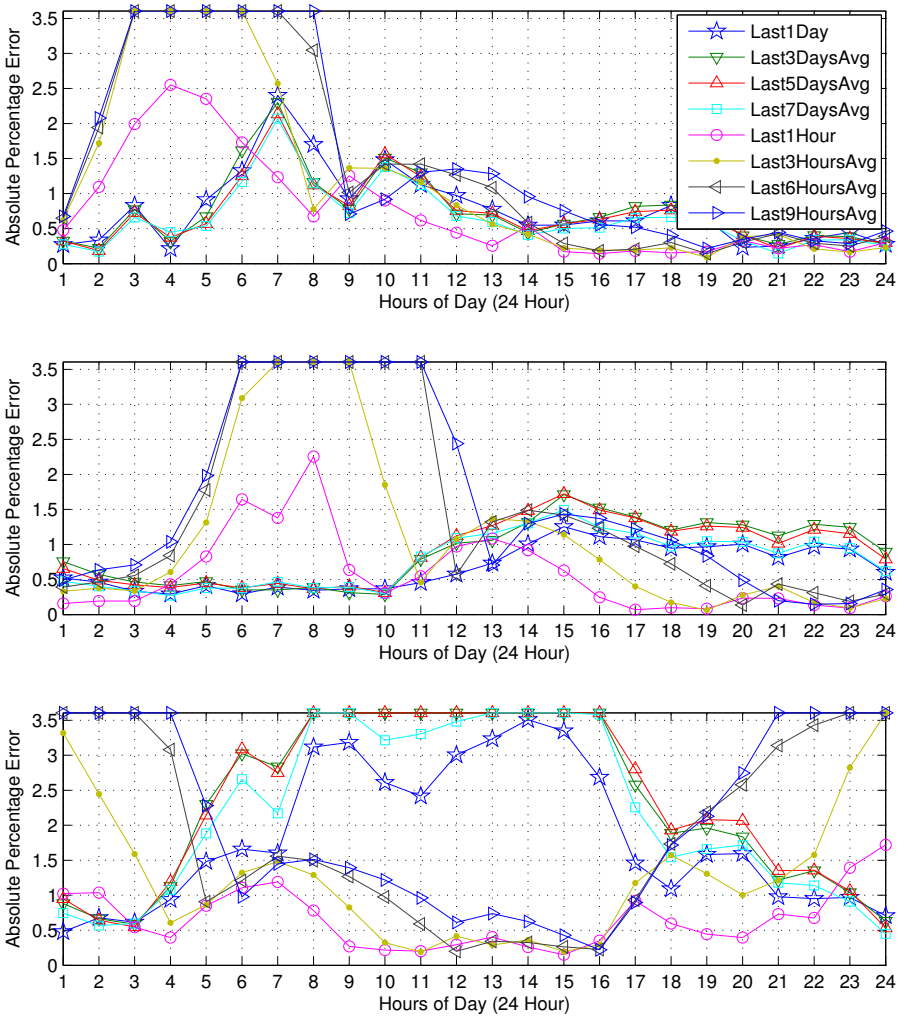

Fig. 4 Predictiveness of different types of past visit volume information across hours of the day for 3 individual properties selected from the Top500Prop. The types of past visit volume information that are more predictive for different hours of the day follow a different pattern across different hours in individual properties (unlike their global behavior) resulting in groups of Web page and time stamp classes with similar visit volume patterns.

class model (i.e., Latent_S_Mod) and the latent time stamp class model (i.e., Latent_T_Mod). The robustness of the latent class model with respect to different number of latent time stamp classes was also measured. The details about the

latent Web page and time-stamp class model is given in Section 2, and the details about the baselines are given in Section 3.2.

Table 4 shows that the proposed probabilistic latent Web page and time stamp class model (i.e., *Latent_ST_Mod*) significantly outperforms (with p-value much less than 0.001) all the baseline approaches as well as the latent Web page class model (Latent_S_Mod) and the latent time stamp class model (Latent_T_Mod) by jointly modeling the latent Web page and time stamp classes that provide much higher modeling flexibility than any of the other models leading to its superior performance across all models. Specifically, it is important to observe that while Laplace_Regr, which does not identify any latent groups, has an (normalized) APE of 0.545 & 0.893 for Top500Prop and Top1000Prop respectively; Latent_$S_{15,20}T_3$_Mod, which identifies latent Web page and time-stamp groups and fits a specialized forecast model for each latent class, achieves an (normalized) APE of 0.453 and 0.798 respectively, which is not only a very significant (p-value much less than 0.0001) performance improvement but also has huge potential impact given the importance of the task (as it directly impacts revenue) in display advertising business. This clearly shows that differentiating the Web pages together with time stamps that jointly share different user visit patterns, and specializing the forecast model for different types/classes of Web pages and time stamps is important for achieving higher forecast accuracy.

It can also be seen from Table 4 that the performance of the Latent_ST_Mod slightly changes with respect to the number of latent Web page classes. Note that the model achieves the best forecast accuracy for the Top500Prop dataset with 15 latent Web page and 3 latent time stamp classes while it achieves the best forecast accuracy for the Top1000Prop dataset with 20 latent Web page and 3 latent time stamp classes. Similar to the observation in Section 4.2, this is due to the fact that a dataset with more properties is more likely to have higher number of Web page and time stamp groups with similar user visit patterns, and therefore more latent Web page and time groups help the model better capture these patterns.

It is important to note that although modeling the latent time stamp classes does not improve the forecast accuracy, modeling latent Web page and time stamp classes jointly achieves the highest forecast accuracy outperforming the latent Web page class model. This is due to the fact that although the predictiveness of different types of past user visit information across hours of day is not significantly different than each other for all properties as shown in Section 4.3, they are different than each other for individual properties. Figure 4 plots the predictiveness of different types of past user visit volume information across hours of day for three different (individual) properties selected from the Top500Prop dataset. It is shown that for different individual properties, importance of different features vary significantly. For instance, it can be seen at the first property (located at the top) that Last1Hour is the most important predictor after 10AM, other features are significantly better for hours between 1AM and 6AM. Similarly at the second property (located in the middle), although Last1Hour is the most important predictor between hours 2PM and 3AM, many other features are much more predictive between hours 4AM and 9AM. An extreme scenario can be observed at the third property (located at the bottom) where Last7DaysAvg, which is one of the most important features for hours between 11PM and 3AM, becomes one of the worst predictors between 8AM and 4PM, while having fluctuating performance

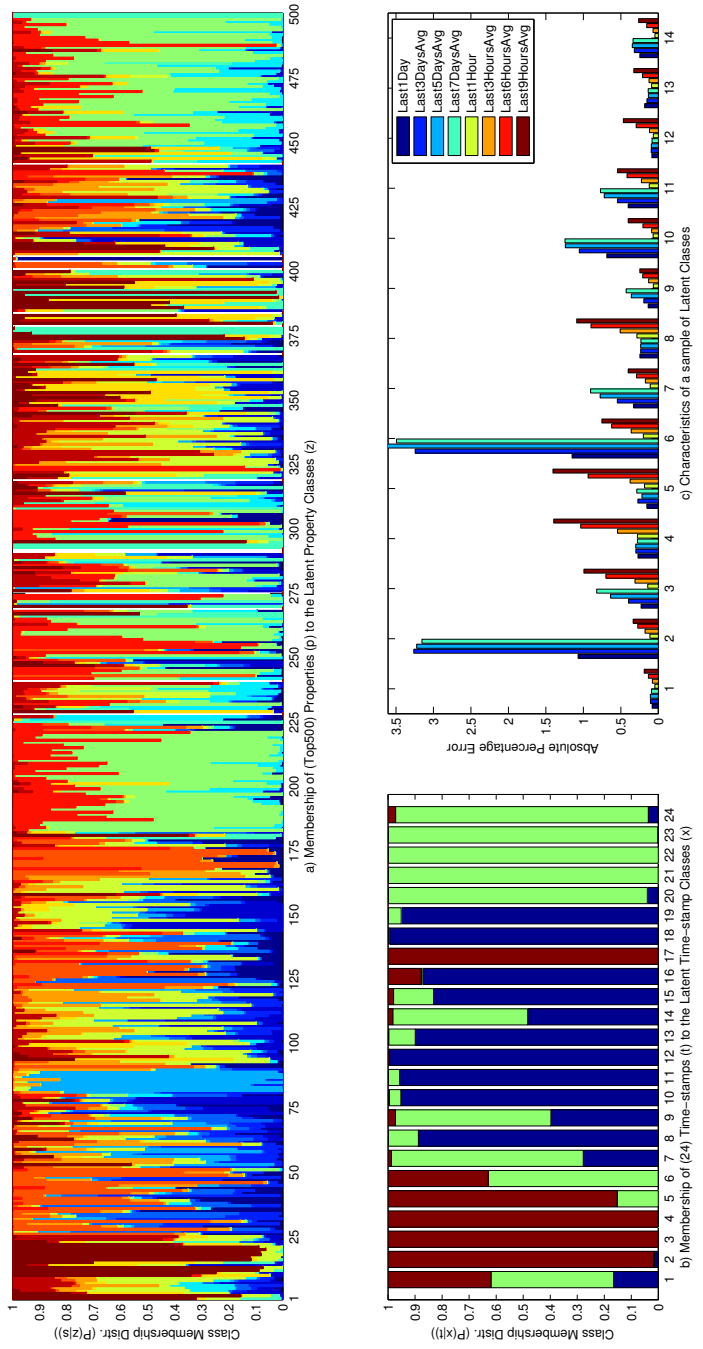

Fig. 5 Memberships of properties to latent property classes in (a), memberships of time-stamps to time-stamp classes in (b), and characteristics of several latent property and time-stamp classes in (c). In subfigures (a) and (b) same colors indicate the membership to the same latent class.

for the rest of the hours. All of those clearly show that during different hours of the day different information types can have very different predictiveness patterns. Therefore it is important to identify types/classes of time stamps (hours) and individual properties that share similar patterns of past user visit information, and learn a specialized model for each property and time stamp class to be able to achieve higher forecast accuracy. The joint latent class model has the capability to capture more fine grained patterns from its much larger space (than the latent Web page class model or the latent time stamp classes model) of property/Web page and time stamp pairs that can be grouped to form the set of joint latent Web page and time stamp classes. Specifically, the joint latent class model has 12K (500 * 24) and 24K (1000 * 24) property and time pairs among which it can look for latent groups while the latent Web page classes model has 500 and 1000 for the Top500Prop and Top1000Prop datasets respectively. In contrast, the latent time stamp classes model has only 24 time stamps to group which is a very limited and much less fine-grained space than the spaces of either of the other latent class models, justifying its weakness in improving the forecast accuracy.

Finally, we analyze the latent Web page and time-stamp groups that have been identified by the proposed probabilistic latent Web page and time stamp class model (i.e., $Latent\_S_{15}T_3\_Mod$). Figure 5-{a,b,c} plots the class memberships of each property to different latent property groups, class memberships of each time-stamp to different latent time-stamp groups, and the characteristics of a sample of identified latent property and time-stamp groups in detail respectively. Specifically, it can be observed from Figure 5-a that properties between ranks 5 & 25, 25 & 50, 80 & 90, 175 and 225, and 450 and 500 have very similar class memberships, which indicates that they have similar user-visit patterns. It is important and very interesting to note that, almost the same pattern can easily be observed from Figure 2, which plots the characteristics of each property in detail. Note that, a large number of properties have quite unique characteristics that are different from the characteristics of other properties, and therefore those properties have weaker memberships to several latent property classes, and can be observed after ranks 225 in Figure 5-a. Memberships of time-stamps to latent time-stamp classes are very interesting as well. It can be observed from Figure 5-b that 24 hours of day are grouped into 3 main groups as hours between 2AM & 5AM, 8AM & 7PM, and 8PM & 12PM, with some transition hours 1AM, 6AM, and 7AM. These latent groups of time-stamps are totally consistent with the common sense, as many properties will have dead hours during the night, other properties will be of more interest during the work hours, and others will be visited more during the after-work hours. It should be noted that such groups are not identified with the latent time-stamp class model, and is only identified with the latent property and time-stamp class model that jointly identifies the latent groups from a larger and more fine-grained space as mentioned before. Last but not the least, Figure 5-c plots the characteristics of some of the identified latent property and time-stamp classes. It can be observed that most of the identified latent classes differ from each other significantly, while some of the latent classes are relatively similar to each other. Note that if the model is forced to utilize a smaller set of latent classes, it will combine those relatively similar latent classes together and generate bigger, yet less fine-grained latent classes, which will eventually deteriorate the performance as observed in Table 4.

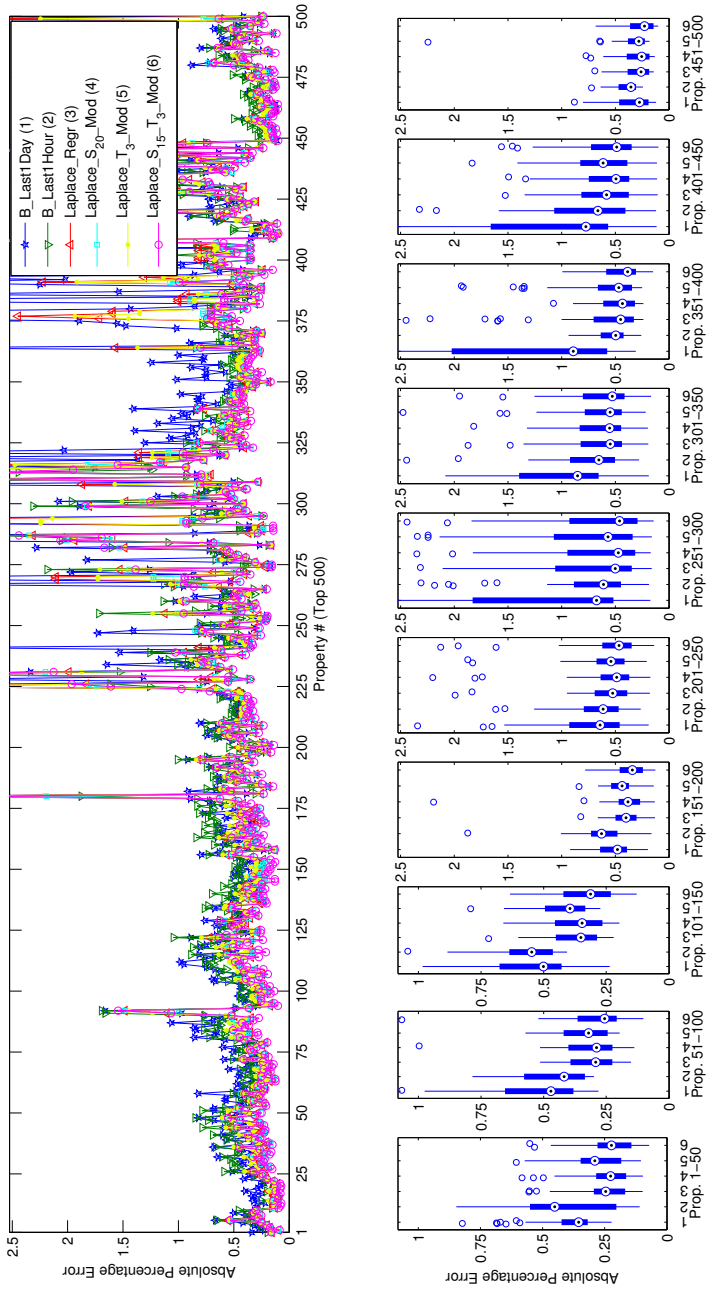

Fig. 6 Performances of the baselines B_Last1Day, B_Last1Hour, Laplace_Regr, and the proposed latent property, latent time-stamp, and latent property and time-stamp class models Latent_S_Mod, Latent_T_Mod, and Latent_S_T_Mod for the top 500 most visited properties in the upper subfigure, and performances of the models (numbered respectively on each column) in detail for groups of properties with different user visit volumes in the lower subfigure. Each of the ten graphs in the lower subfigure is a boxplot summarizing the performances of the models on each column.

4.5 Analysis of the Robustness of the Baselines and Proposed Models

This section analyzes the performances of the baselines B_Last1Day, B_Last1Hour, and Laplace_Regr as well as the proposed latent property class model Latent_S_Mod, latent time-stamp class model Latent_T_Mod, and latent property and time-stamp class model Latent_S_T_Mod in detail, focusing on strengths and weaknesses of each model.

Specifically, Figure 6 plots performances of the models for the top 500 most visited properties. It is important to observe that all models perform relatively better for the most visited properties. Especially for the properties before rank 225 and after rank 450, all properties have APE values less than 0.5, whereas for properties between ranks 225 & 450 the APE values are clearly higher. This can be explained by the fact that the most visited properties have so many user visits that small fluctuations in users' visits does not change the overall user visit pattern of the properties, which eventually makes the overall user visit patterns easier to forecast. Similarly, the user visits for properties with ranks after 450 start having smaller number of user visits such that there are less number of users causing a fluctuation in the overall user visits. Yet, the properties with ranks between 225 & 450 can be seen as a transition set of properties that have enough number of users to result in significant fluctuations in the overall user visits, and yet does not have enough number of users to alleviate the effect of those fluctuations in the overall user visit patterns. Therefore these properties have much more complex user visit patterns, which makes the forecasting task much harder. Indeed, it is very interesting to note that this can also be observed in Figure 5 that the properties between ranks 225 & 450 have very complex memberships to the latent property classes, which clearly indicates that they have very complex user visit patterns, and can not be explained (associated) with a small number of latent property classes.

Figure 6 also plots the performance comparison of the models for properties with different user visit volumes in detail (with the lower subfigure). It can be seen that while B_Last1Day and B_Last1Hour compete with each other for different groups of properties, B_Last1Day mostly outperforms B_Last1Hour for the properties with more stable user visit patterns (with ranks less than 250 and ranks after 450 as mentioned above). On the other hand, for properties with more complex user visit patterns (i.e., with ranks between 250 & 450), B_Last1Hour is observed to perform better. This shows that when there is significant fluctuation in the user visits, it is helpful to utilize the most recent data to achieve more effective results. It is important to observe that the performances of both B_Last1Day and B_Last1Hour methods are not robust across different properties, and are therefore not reliable. For instance, the performance of B_Last1Day fluctuates significantly for properties between ranks 250 & 450 while the performance of B_Last1Hour fluctuates significantly for properties between ranks 1 & 100, and 401 & 450. Therefore both methods are not reliable enough to be used as a forecasting method for all properties. On the other hand, Laplace_Regr outperforms both B_Last1Day and B_Last1Hour significantly for all groups of properties, and has much less variance in its performance, achieving much more effective and reliable performance. Finally, the proposed models have significantly better performance than the baselines, and have much less variance. Specifically, Latent_S_T_Mod is not only the best performing model that significantly outperforms all other models

for all groups of properties with different user visit volumes, but also has one of the lowest variances in performance for all groups of properties. This clearly shows that Latent_S_T_Mod is not only the best performing model overall, but also is the best performing model for properties with different characteristics/visit volumes, and therefore is the most robust and reliable model.

5 Conclusions

Online advertising is one of the most profitable business models for Internet services that is growing very fast as consumers shift their time to digital media. Online display advertising is one of the major types of online advertising, where advertisers buy targeted user visits from publishers in order to promote their products by displaying graphical (e.g. image, video) advertisements on popular Web pages. An important problem in online display advertising is how to forecast the number of user visits for a Web page during a particular period of time. Different Web pages and different time stamps have different user visit trends, and it is important to learn specialized forecasting models for properties/Web pages and time stamps with different user visit trends. Prior research addressed the problem by using traditional time-series forecasting techniques on historical data of user visits by fitting a single regression model built for forecasting based on historical data for all Web pages, and did not fully explore the fact that different types of Web pages and different time stamps have different patterns of user visits.

In this paper, we propose a series of probabilistic latent class models that automatically identifies latent classes for Web pages and time stamps with similar user visit trends, and learns a separate forecasting model for each type/class of Web pages and time stamps. It is shown that the proposed probabilistic latent Web page and time stamp class model achieves much better modeling flexibility by being able to differentiate the importance of different types of information across different Web page and time stamp classes. An efficient learning algorithm based on Expectation and Maximization algorithm has been proposed to simultaneously learn the model parameters. Thorough empirical studies have been conducted with a real-world dataset from Yahoo!. Experimental results and detailed analysis demonstrate the effectiveness and robustness of the proposed probabilistic latent class models for forecasting user visits in online display advertising.

There are several possibilities to extend the research. Users with different visiting habits are not differentiated with the current model. Future work will focus on how to model users with similar visiting behaviors, and how to integrate the current work with the user model into a unified framework.

Acknowledgements This research was partially supported by the National Science Foundation research grants IIS-0746830, CNS-1012208, IIS-1017837, and a research grant from Yahoo!. Any opinions, findings, conclusions, or recommendations expressed in this paper are the authors', and do not necessarily reflect those of the sponsor.

References

Agarwal D, Broder AZ, Chakrabarti D, Diklic D, Josifovski V, Sayyadian M (2007) Estimating rates of rare events at multiple resolutions. In: Proceedings

of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '07, pp 16–25

Agarwal D, Gabrilovich E, Hall R, Josifovski V, Khanna R (2009) Translating relevance scores to probabilities for contextual advertising. In: Proceeding of the 18th ACM conference on Information and knowledge management, ACM, New York, NY, USA, CIKM '09, pp 1899–1902

Agarwal D, Agrawal R, Khanna R, Kota N (2010a) Estimating rates of rare events with multiple hierarchies through scalable log-linear models. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '10, pp 213–222

Agarwal D, Chen D, Lin Lj, Shanmugasundaram J, Vee E (2010b) Forecasting high-dimensional data. In: Proceedings of the 2010 International ACM SIGMOD Conference on Management of Data, ACM, New York, NY, USA, SIGMOD '10, pp 1003–1012

Aggarwal G, Goel A, Motwani R (2006) Truthful auctions for pricing search keywords. In: Proceedings of the 7th ACM Conference on Electronic Commerce, ACM, New York, NY, USA, EC '06, pp 1–7

Alaei S, Arcaute E, Khuller S, Ma W, Malekian A, Tomlin J (2009) Online allocation of display advertisements subject to advanced sales contracts. In: Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising, ACM, New York, NY, USA, ADKDD '09, pp 69–77

Bharadwaj V, Ma W, Schwarz M, Shanmugasundaram J, Vee E, Xie J, Yang J (2010) Pricing guaranteed contracts in online display advertising. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, CIKM '10, pp 399–408

Bishop CM (2006) Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA

Boyd S, Vandenberghe L (2004) Convex Optimization. Cambridge University Press, New York, NY, USA

Broder A, Fontoura M, Josifovski V, Riedel L (2007) A semantic approach to contextual advertising. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '07, pp 559–566

Cetintas S, Chen D, Si L, Shen B, Datbayev Z (2011a) Forecasting counts of user visits for online display advertising with probabilistic latent class models. In: Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '11, pp 1217–1218

Cetintas S, Rogati M, Si L, Fang Y (2011b) Identifying similar people in professional social networks with discriminative probabilistic models. In: Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '11, pp 1209–1210

Chakrabarti D, Agarwal D, Josifovski V (2008) Contextual advertising by combining relevance with click feedback. In: Proceeding of the 17th International Conference on World Wide Web, ACM, New York, NY, USA, WWW '08, pp 417–426

Cui Y, Zhang R, Li W, Mao J (2011) Bid landscape forecasting in online ad exchange marketplace. In: Proceedings of the 17th ACM SIGKDD International

Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '11, pp 265–273

Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society

Fang Y, Si L, Mathur AP (2011) Discriminative probabilistic models for expert search in heterogeneous information sources. Inf Retr 14:158–177

Feige U, Immorlica N, Mirrokni V, Nazerzadeh H (2008) A combinatorial allocation mechanism with penalties for banner advertising. In: Proceedings of the 17th International Conference on World Wide Web, ACM, New York, NY, USA, WWW '08, pp 169–178

Hatch A, Bagherjeiran A, Ratnaparkhi A (2010) Clickable terms for contextual advertising. In: Proceedings of the Fourth International Workshop on Data Mining and Audience Intelligence for Advertising, ACM, New York, NY, USA, ADKDD '09, pp 69–77

IAB, PricewaterhouseCoopers (2011) Iab internet advertising revenue report

Karimzadehgan M, Li W, Zhang R, Mao J (2011) A stochastic learning-to-rank algorithm and its application to contextual advertising. In: Proceedings of the 20th International Conference on World Wide Web, ACM, New York, NY, USA, WWW '11, pp 377–386

Labrou Y, Finin T (1999) Yahoo! as an ontology: using yahoo! categories to describe documents. In: Proceedings of the 8th International Conference on Information and Knowledge Management, ACM, New York, NY, USA, CIKM '99, pp 180–187

Lacerda A, Cristo M, Gonçalves MA, Fan W, Ziviani N, Ribeiro-Neto B (2006) Learning to advertise. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '06, pp 549–556

Lahaie S, Parkes DC, Pennock DM (2008) An expressive auction design for online display advertising. In: Proceedings of the 23rd National Conference on Artificial Intelligence, AAAI Press, pp 108–113

Murdock V, Ciaramita M, Plachouras V (2007) A noisy-channel approach to contextual advertising. In: Proceedings of the 1st International Workshop on Data mining and Audience Intelligence for Advertising, ACM, New York, NY, USA, ADKDD '07, pp 21–27

Ribeiro-Neto B, Cristo M, Golgher PB, Silva de Moura E (2005) Impedance coupling in content-targeted advertising. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '05, pp 496–503

Richardson M, Dominowska E, Ragno R (2007) Predicting clicks: estimating the click-through rate for new ads. In: Proceedings of the 16th International Conference on World Wide Web, ACM, New York, NY, USA, WWW '07, pp 521–530

Shumway RH, Stoffer DS (2007) Time Series Analysis and Its Applications. Springer

Wang X, Broder A, Fontoura M, Josifovski V (2009) A search-based method for forecasting ad impression in contextual advertising. In: Proceedings of the 18th International Conference on World wide web, ACM, New York, NY, USA, WWW '09, pp 491–500

Yan R, Hauptmann AG (2006) Probabilistic latent query analysis for combining multiple retrieval sources. In: Proceedings of the 29th Annual International

ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '06, pp 324–331

Yang J, Vee E, Vassilvitskii S, Tomlin J, Shanmugasundaram J, Anastasakos T, Kennedy O (2010) Inventory allocation for online graphical display advertising. Arxiv preprint arXiv:10083551

Zellner A, Tobias J (1999) A note on aggregation, disaggregation and forecasting performance. Journal of Forecasting