Personalization from Incomplete Data: What You Don't Know Can Hurt

Balaji Padmanabhan, Zhiqiang Zheng and Steven O. Kimbrough
The Wharton School, University of Pennsylvania
3620 Locust Walk, Philadelphia, PA 19104-6366
tel:+1(215)573-9646, fax:+1(215)898-3664
{balaji, zhengzhi, kimbrough}@wharton.upenn.edu

ABSTRACT
Clickstream data collected at any web site (site-centric data) is inherently incomplete, since it does not capture users' browsing behavior across sites (user-centric data). Hence, models learned from such data may be subject to limitations, the nature of which has not been well studied. Understanding the limitations is particularly important since most current personalization techniques are based on site-centric data only. In this paper, we empirically examine the implications of learning from incomplete data in the context of two specific problems: (a) predicting if the remainder of any given session will result in a purchase and (b) predicting if a given user will make a purchase at any future session. For each of these problems we present new algorithms for fast and accurate data preprocessing of clickstream data. Based on a comprehensive experiment on user-level clickstream data gathered from 20,000 users' browsing behavior, we demonstrate that models built on user-centric data outperform models built on site-centric data for both prediction tasks.

Categories and Subject Descriptors
H.2.8 **[Database Management]**: Applications – *Data Mining*; I.2.6 **[Artificial Intelligence]**: Learning

Keywords
Incomplete data, learning, personalization, clickstream data, data preprocessing, probabilistic clipping

1. INTRODUCTION
The literature on mining clickstream data collected by a site is replete with examples illustrating the power of using such data for various personalization applications such as optimizing web site design dynamically based on navigational patterns [20,21,25], predicting the likelihood of a purchase at a site [11,18,26] and designing effective one-to-one marketing strategies [1,2,5,12,13]. A characteristic of most of the existing approaches to personalization is that these methods build profiles and models based on data collected *by a single web site* about users' accesses to its site. We refer to such data as *site-centric* data, which we define to be clickstream data collected at a site augmented with user demographics and cookies to identify users [24][1]. In a sense, traditional approaches using site-centric are *myopic* – they are based on firms building models from data collected at their site only. However, the myopic nature of most current personalization methods is *not* due to the fact that site-centric data is *adequate* for understanding customer behavior; rather, it is due to the nature of data ownership constraints – most sites only have access to their own logfiles.

For example, consider two users who browse the web for air tickets. Assume that the first user's session is as follows $Cheaptickets_1$, $Cheaptickets_2$, $Travelocity_1$, $Travelocity_2$, $Expedia_1$, $Expedia_2$, $Travelocity_3$, $Travelocity_4$, $Expedia_3$, $Cheaptickets_3$ where X_i represents some page i, at website X. In this session assume that the user purchases a ticket at Cheaptickets. Further assume that the second user's session is $Expedia_1$, $Expedia_2$, $Expedia_3$,, $Expedia_4$ and that this user purchases a ticket at Expedia (in the booking page $Expedia_4$, in particular). Expedia's (site-centric) data would include the following:

User1: $Expedia_1$, $Expedia_2$, $Expedia_3$
User2: $Expedia_1$, $Expedia_2$, $Expedia_3$, $Expedia_4$

In one case (user 2) the first three pages result in the user booking a ticket at the next page. In the other case (user 1), the first three pages result in no booking. Expedia sees the "same" initial browsing behavior, but with opposite results – one which resulted in a booking and one which did not. The problem confronting Expedia here is the inherent incompleteness of data that the site collects – user1's browsing behavior is not completely known to Expedia.

It is easy to see that however sophisticated the personalization algorithms used, it is almost impossible to differentiate these two sessions based on site-centric data alone. Nevertheless, most conventional personalization techniques are asked to do so, despite the lack of complete information. Hence given the incomplete nature of site-centric data, can personalization models actually 'work' on such data?

This paper presents initial results from studying the quantitative and qualitative impacts of learning from incomplete data. To do so, we construct a "complete" version of site-centric data as

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
KDD 01 San Francisco CA USA
Copyright ACM 2001 1-58113-391-x /01/08...$5.00

[1] Site-centric data is similar to traditional web logfile data. We make a distinction since we permit site-centric data to include additional user-level information, such as demographics, that a site can collect.

follows. We define *user-centric* data to be site-centric data *plus* data on where else the user went in the current session. In the above example, the user centric data for Expedia will be:

User1:$Cheaptickets_1$, $Cheaptickets_2$, $Travelocity_1$, $Travelocity_2$, $Expedia_1$, $Expedia_2$, $Travelocity_3$, $Travelocity_4$, $Expedia_3$, $Cheaptickets_3$

User2: $Expedia_1$, $Expedia_2$, $Expedia_3$, $Expedia_4$

It is important to note that the "complete" data scenario is hypothetical, *but in reality is the complete scenario*. For most sites, it is impossible for the site to keep track of a user's browsing behavior at other sites. Nevertheless, we believe that this comparison is essential since it provides a benchmark against which approaches using site-centric data can be evaluated.

In this paper, we compare models derived from site-centric data to those derived from user-centric data. Such a comparison is hard in general, since the space of potential models is not enumerable. Hence we consider two classes of models:

Session-level prediction. In this class we constructed models in prior work [18] that predict whether the remainder of a current user's session will result in a purchase. In this paper we extend our prior work in two directions: (i) we present in detail new algorithms for correct preprocessing of clickstream data and (ii) we consider an additional class of models for user-level prediction.

User-level prediction: In this class we construct models that predict whether a given user at a given point in time will make a purchase at the site during some future session.

Within each of these two classes of models, we consider four classifiers - linear regressions, logistic regressions, classification trees and neural networks. Each classifier is then built on both site-centric and user-centric data. In order to build these classifiers correctly, site-centric and user-centric data need to be preprocessed carefully. As mentioned above, we present new algorithms for these preprocessing tasks. The resulting eight pairs of classification models are then compared quantitatively based on lift and qualitatively based on model interpretation.

The main results of this paper are:

1. The models built on user-centric data outperformed models built on site-centric data *in all the eight comparisons*. This result is robust across four different classifiers that span the spectrum from linear to log-linear and non-linear classifiers. Though the nature of the result is not surprising, the magnitude of the difference between site-centric and user-centric approaches is strikingly high (almost a factor of two in some cases), which indicates that the cost of incomplete information may be much higher than is currently expected.
2. The result is more striking for session-level prediction than for user-level prediction. This finding suggests that the effect of learning from incomplete data varies based on the personalization task.
3. Examples of qualitative findings indicate that it is possible to make potentially erroneous conclusions based on site-centric data alone.

These results provide evidence of the potential pitfalls of current personalization methods and suggest the need for additional evaluation and further research.

The rest of this paper is organized as follows. In Section 2 we present an overview of prior work. In Section 3 we describe the data and discuss how to generate site-centric and user-centric data. Section 4 discusses the problem formulation for the session-level prediction task based on site-centric and user-centric data and presents a new algorithm for preprocessing clickstream data applicable for session-level prediction problems. The problem formulation for the user-level prediction task on site-centric and user-centric data and the applicable data preprocessing algorithm is presented in Section 5. Results of solving these two problems based on four different classification methods are presented in Section 6. Finally section 7 presents conclusions.

2. PRIOR WORK

Other than the session-level prediction problem studied in our prior work [18], there is little prior work in comparing models built on incomplete data to those built on complete data in the context of web usage mining. In [18] we built classifiers to predict the session-level prediction problem in the context of site-centric and user-centric data. The new contributions in this paper are (1) new preprocessing algorithms for fast and accurate preprocessing of clickstream data and (2) considering an additional prediction task and comparing the results across these tasks.

The literature on personalization models built from site-centric data is vast and spans mainstream business research and work in data mining. In the Information Systems and Marketing literature, several models have been proposed [14,15,24] to study which user visits at a web site actually lead to purchases. Moe & Fader [14] use web usage data to predict a customer's probability of purchasing at any given visit based on prior visits and purchases. Their results indicate that a consumer's history and purchasing threshold are highly predictive of purchasing propensity in a given session. The study was based on usage data from Amazon.com. In a subsequent study, Moe & Fader [15] study visits at two online stores (Amazon and CDNOW) separately and found that consumers' searching behavior evolves with accumulated experience. Sen et al. [24] study the information needs of marketers and provide a framework for understanding how much of these needs can be satisfied from clickstream data collected at a web site. There has been prior work in studying models built on user accesses across multiple sites such as [7] and [19]. Though these studies indicate the need to examine usage across sites, they do not study the problem of comparing models built on site-centric and user-centric data.

In the data mining community, building user profiles based on transactional history has been studied in [1,2,5,12,25,17,23]. Theusinger and Huber analyzed users' navigation histories to optimize web site design [25]. Nasraoui et al. proposed clustering user sessions to predict future user behavior by mining the usage gleaned from the site of their computer science department [17]. Schechter et al. developed techniques for using path profiles of users to predict future web page requests [23]. In the context of building user profiles from transaction data, prior work described alternate approaches to building and evaluating user profiles based on transaction history [1,2,5,12].

In this section we presented an overview of prior work. In the next section we describe how to construct site-centric and user-centric data from commercially available user-level browsing data.

3. DATA

In this section we first describe the data used in these experiments and then describe how to approximate site-centric and user-centric data. The raw data provided to us from a market data vendor consisted of records of 20,000 users' web surfing behavior over a period of 6 months (userIDs were anonymized to mask identities). The users were chosen based on standard market survey techniques and are assumed to represent a random sample of US households. The data included user demographics and transaction history over the entire period. The total size of the raw data amounted to 30GB and represented approximately 4 million user sessions. In addition, the vendor manually categorized the various sites accessed by the users into categories such as books, search engines, news, CDs, travel etc. In particular, since we focus on predicting whether a session will result in a *booking*, we chose five categories among them (book, music, travel, auction and general shopping mall) that represent sites that sell products. The number of user sessions containing one of these categories was 310,323. Firms such as MediaMetrix and Netratings collect such data.

Site-centric data can be constructed by taking each user session from this data and constructing snapshots for each unique site in the session such that the snapshot consists of pages belonging to that particular site. For example, given a single session $<Cheaptickets_1, Cheaptickets_2, Travelocity_1, Travelocity_2, Expedia_1, Expedia_2, Travelocity_3, Travelocity_4, Expedia_3, Cheaptickets_3>$, the records in the site-centric data contain:

1. $<Cheaptickets_1, Cheaptickets_2, Cheaptickets_3>$ for site *Cheaptickets*
2. $<Travelocity_1, Travelocity_2, Travelocity_3, Travelocity_4>$ for site *Travelocity*, and
3. $<Expedia_1, Expedia_2, Expedia_3>$ for site *Expedia*.

In general, each record in site-centric data contains user demographics, the set of pages visited in a session at a given site and the site ID. The number of records in the site-centric data constructed in this manner is the total number of user sessions across all the sites among the shopping category sites considered in this paper.

Observe that in effect the procedure outlined above "recreates" each site's partial logfile data based on the sample of 20,000 users' web accesses. Given that it is impossible to obtain the complete logfile data for every commercial site, we believe that the strength of this procedure is that it can simulate individual logfiles from user-level browsing data. The limitation is that in our study we only have 20,000 users' browsing behavior. However, we have no reason to believe that this is not a representative sample - the market data vendor's data gathering methodology is based on standard market research techniques for ensuring a random sample.

The creation of user-centric data for each site is straightforward. For each site s, the subset of all user sessions that contain s represent the user-centric data for the site. For instance, the user session $<Cheaptickets_1, Cheaptickets_2, Travelocity_1, Travelocity_2, Expedia_1, Expedia_2, Travelocity_3, Travelocity_4, Expedia_3, Cheaptickets_3>$ will belong to the user-centric data of Expedia, Travelocity and Cheaptickets.

In general, each record in user-centric data contains user demographics, the complete set of pages visited in a session and the name of the site. The number of records in the user-centric data constructed in this manner is the same as the number of records in the site-centric data – the only difference being the set of pages represents complete information regarding where else a user visited during this session.

In this section we described how to construct site-centric and user-centric data from commercially available user-level browsing behavior data. In the next section we describe the session-level prediction problem and present a new algorithm for accurate preprocessing of clickstream data for session-level prediction tasks.

4. SESSION-LEVEL PREDICTION

The session-level prediction problem is predicting at any given point in a user's session at a web site if the remainder of the session will result in a booking at the site. To illustrate the difference in using site-centric data and user-centric data consider the example of User1's session at Expedia (from Section 1). Models built on site-centric data, for example, make a prediction based on the session fragment $Expedia_1, Expedia_2, Expedia_3$ while models built on user-centric data make this prediction based on the session fragment $Cheaptickets_1, Cheaptickets_2, Travelocity_1, Travelocity_2, Expedia_1, Expedia_2, Travelocity_3, Travelocity_4, Expedia_3$. In this section we first present a new algorithm for preprocessing site-centric data appropriately for this task. We then present a brief overview of preprocessing user-centric data for this task (since it is similar to the previous method except for one key difference).

4.1 Site-Centric Data Preprocessing

Given that the goal is to predict at any point if the session will result in a booking, the training and testing data for the model should come from prior cases of user known sessions, some of which resulted in booking and some of which may not have. Consider a specific user's session at a web site $<p_1, p_2, p_3, p_4, p_5>$. Assume that the user made a booking in page p_4. This single session is *not* a single data record for modeling. Rather, it provides 5 data records:

1. A session that began with p_1 resulted in the user booking at a subsequent point.
2. A session that began with p_1, p_2 resulted in booking at a subsequent point.
3. A session that began with p_1, p_2, p_3 resulted in booking at a subsequent point.
4. A session that began with p_1, p_2, p_3, p_4 did *not* result in booking at a subsequent point.
5. A session that began with p_1, p_2, p_3, p_4, p_5 did *not* result in booking at a subsequent point.

This distinction is important. It indicates that sessions create data records proportional to their length. Assuming that every additional page accessed provides additional information, this is an appropriate method to preprocess user sessions for the session-level prediction problem. In general, a session of length k provides k data records for modeling. In total, the number of records is therefore the sum of all session lengths (total number of pages accessed). Given the potential explosion in the number of

data points this creates, rather than explicitly creating k data records for each session of length k we sample each session probabilistically based on its length. For example, if a sampling rate is 0.2, a session of length k on average provides $0.2*k$ data records for modeling.

Associated with each sampling, a random 'clipping point' is chosen within the session; information before the clipping point is used to predict whether a purchase occurs in the part of the session after the clipping point. The data associated with this sample therefore consists of all pages before the clipping point and an indicator variable that represents whether a purchase occurred in the fraction after the clipping point[2]. This indicator variable is the dependent variable used in the predictive models. The explanatory variables are constructed based on user-defined functions that summarize data on this user based on all information known before the prediction point.

A key strength of probabilistic sampling is that the size of the preprocessed data can be chosen based on time and space constraints available. Choosing a maximum desired data size, *dnum*, is equivalent to choosing a sampling rate per session of *dnum/numtotal*, where *numtotal* is the sum of the lengths of all the sessions in the data. This is a major advantage since otherwise, the size of the data is the number of user clicks at a web site. For most sites, this is an unmanageable number to feed into a classifier, given space and time constraints. In the rest of this section, we present *ProbabilisticClipping*, an algorithm for preprocessing site-centric data for session-level prediction task. Before presenting the algorithm, we first present some preliminaries.

Let S_1, S_2, …, S_N be N user sessions in a site's site-centric data. Assume that in this data the number of unique users is M and users are identified by a *userid* $\in \{1,2,\ldots,M\}$. We define each session S_i to be a tuple of the form $<u_i, C_i>$ where u_i is the userid corresponding to the user in session S_i and C_i is a set of tuples of the form *<page, accessdetails>*, where each tuple represents data that a site captures on each user click. Corresponding to a click, *page* is the page accessed and *accessdetails* is a set of attribute-value pairs that represents any other information that a site can capture from each user click. This includes standard information from http headers such as time of access, IP address, referrer field etc *and* other information such as whether the user made a purchase in this page. In particular we assume that *accessdetails* necessarily contains information on the time a page is accessed. For example, based on the above representation scheme, three user sessions at Expedia are represented as follows:

S_1 = <1, {<home.html, { (time, 02/01/2001 23:43:15), (IP, 128.122.195.3) } >,
<flights.html, { (time, 02/01/2001 23:45:15), (IP, 128.122.195.3) } >,
<hotels.html, { (time, 02/01/2001 23:45:45), (IP, 128.122.195.3) } >
} >

S_2 = <2, {<home.html, { (time, 02/01/2001 23:50:10), (IP, 128.122.197.23) } >,
<cars.html, { (time, 02/01/2001 23:55:15), (IP, 128.122.197.23) } >
} >

S_3 = <1, {<home.html, { (time, 02/02/2001 07:43:07), (IP, 128.122.195.3) } >,
<cars.html, { (time, 02/02/2001 07:45:49), (IP, 128.122.195.3) } >
} >

Given a session $S_i = <u_i, C_i>$, define function *kth-click*(S_i, j) that returns a tuple *<page, accessdetails>* $\in C_i$ if *page* is the j^{th} page accessed in the session as determined from the time each page is accessed. In the above example *kth-click*(S_1, 2) is <flights.html, { (time, 02/01/2001 23:45:15), (IP, 128.122.195.3) } >.

Also we define function *fragment*(S_i, j, fraglength) that represents data captured from a set of consecutive clicks in the session. In particular, *fragment*(S_i, j, fraglength) is the set of all tuples *kth-click*(S_i, m) such that $j \le m \le$ minimum((j + fraglength −1), $|C_i|$), where $S_i=<u_i,C_i>$. For example, *fragment*(S_1,2,2)= { <flights.html, { (time, 02/01/2001 23:45:15), (IP, 128.122.195.3) }>, <hotels.html, { (time, 02/01/2001 23:45:45), (IP, 128.122.195.3) } > }. For any given set, f, of *<page, accessdetails>* pairs, and any given session S_i, we say f is a fragment of S_i if there exists j, k such that f = *fragment*(S_i , j, k).

Finally, much prior work [8,11,13,26] in building online customer interaction models assumes that three sets of variables are particularly relevant:

1. Current visit summaries (e.g. time spent in current session).
2. Historical summaries of the user (e.g. average time spent per session in the past).
3. User demographics.

Since the specific variables created in these three categories can be application specific, we assume three user-defined functions as inputs to the algorithms:

1. summarize_current(f, S_i), defined when f is a fragment of S_i. This function is assumed to return user-defined summary variables for the current fragment and session. For example for the running example used in this section, summarize_current(*fragment*(S_1, 1, 2), S_1) may return *numpages=2, tot_time=150 seconds, booked = 1* assuming the user made a booking in one of the three pages accessed in the session.
2. summarize_historical(f, S, i), where f is a fragment of session S_i and S = {S_1, S_2,…, S_N}. This function is assumed to return summary variables based on all previous sessions. Note that the historical summaries are usually about the specific user in session S_i.
3. demographics(u_i) which returns the demographic information available about user u_i.

[2] We use a heuristic that considers properties of secure-mode transactions to infer bookings. In prior research [18] we describe the heuristic and show that it is reasonable and necessary.

```
Inputs: (a) User sessions S1, S2, …, SN
        (b) Desired number of data records, dnum
        (c) functions summarize_current, summarize_historical, demographics

Outputs: Data records D1, D2,…, DP.

1   S = S1 U S2… U SN
2   numtotal = 0
3   for (i = 1 to N) {
4        <u, C> = Si
5        numtotal = numtotal + |C|
6   }
7   samplerate = dnum/numtotal
8   p = 0; i = 1
9   while (p < dnum) {
10       <u, C> = Si
11       session_len = |C|
12       rand = random_real(0,1)
13       if (rand < session_len * samplerate) { /* whether to sample */
14            clip = random_int(1, session_len) /* which point to clip */
15            f = fragment(Si, 1, clip)
16            current = summarize_current(f, Si)
17            history = summarize_historical(f, S, i)
18            demog = demographics(u)
19            DP = current U history U demog
20            p = p + 1
21            output 'DP'
22       }
23   i = i + 1
24   if (i > N) {i = 1}
25  }
```

Figure 4.1 Algorithm Probabilistic Clipping

Most prior work in the marketing and the data mining literature uses different set of usage metrics for different purposes [6,7,9, 10,16,22]. Borrowing on much of this prior work, for the experiments presented in this paper, we define *summarize_historical* and *summarize_current* to return 9 metrics for site-centric data. In addition, we use 6 demographic variables and 1 variable indicating the category of the site. In total, the site-centric data contains 16 (7+9) explanatory variables. These usage metrics are presented in detail in [18].

Given these preliminaries, we now present ProbabilisticClipping, an algorithm for preprocessing site-centric data for session-level prediction task. The inputs to the algorithm are:

1. A set of user sessions at a site $S_1, S_2, \ldots, S_N$.
2. Desired number of data records, *dnum*.
3. Functions *summarize_current, summarize_historical, demographics*

The output are processed data records $D_1, D_2, \ldots, D_P$.

The algorithm is presented in Figure 4.1. Based on the desired data size, the sample rate is first computed (as described previously) in steps 1-7. Steps 9-25 iterates over all the sessions repeatedly until the desired number of records is sampled. Each time, a session is sampled probabilistically based on the expected number of records that should be derived from it.

4.2 User-Centric Data Preprocessing

In this section we briefly present the main ideas behind preprocessing user-centric data for the session-level prediction problem. This involves using the Probabilistic Clipping method partially – only for creating fragments of all sessions in site-centric data and then augmenting these fragments with user-centric information before create the summary variables. For example, consider a single user's session. Let the records in the site-centric and user-centric data for Expedia contain the following records:

site-centric record: $Expedia_1$, $Expedia_2$, $Expedia_3$
user-centric record: $Cheaptickets_1$, $Cheaptickets_2$, $Travelocity_1$, $Travelocity_2$, $Expedia_1$, $Expedia_2$, $Travelocity_3$, $Travelocity_4$, $Expedia_3$

If the sampling rate is 0.7 assume that as part of the Probabilistic Clipping procedure for site-centric data, the following 2 fragments of the session are created $Expedia_1$ and $Expedia_1$, $Expedia_2$, $Expedia_3$. Preprocessing user-centric data for Expedia would then create the fragments $Cheaptickets_1$, $Cheaptickets_2$, $Travelocity_1$, $Travelocity_2$, $Expedia_1$ and $Cheaptickets_1$, $Cheaptickets_2$, $Travelocity_1$, $Travelocity_2$, $Expedia_1$, $Expedia_2$, $Travelocity_3$, $Travelocity_4$, $Expedia_3$. Based on these two fragments, summary variables are created similar to what was done before.

```
Inputs: (a) User sessions S1, S2, …, SN
        (c) functions summarize_historical, demographics

Outputs: Data records D1, D2,…, DN.

1   S = S1 U S2… U SN
2   p = 0
3   for (i = 1 to N) {
4         <u, C> = Si
5         session_len = |C|
6         history = summarize_historical(S, i)
7         demog = demographics(u)
8         Di = history U demog
9         output 'Di'
10  }
```

Figure 5.1 Algorithm UserLevelDP

Clearly the summary variables in user-centric data will contain additional metrics such as percentage of total hits to Expedia's site in the current session (20% for the first fragment and 33% for the second fragment in the example). In addition to the 9 metrics derived for site-centric data, we define *summarize_historical* and *summarize_current* to return 25 additional metrics for user-centric data – the difference is because user-centric data just has much more information. We also use 6 demographic variables and 1 variable indicating the category of the site – these additional 7 variables are common to both site and user-centric data. In total, the site-centric data contains 16 (7+9) explanatory variables while the user-centric data contains 41 (7+9+25). A detailed set of user-centric metrics is presented in [18].

In this section we described the session-level prediction problem and presented a new algorithm, Probabilistic Clipping, for accurate preprocessing of clickstream data for session-level prediction tasks for site-centric data. We also presented a brief overview of the preprocessing method for the same task in the case of user-centric data. In the next section we describe the second problem based on which we compare site-centric and user-centric data – the "user level" prediction problem[3].

5. USER-LEVEL PREDICTION

The session-level prediction problem was predicting at any given point *within* a user's session at a web site if the remainder of the session will result in a booking at the site. The user-level prediction problem is predicting after any given session, whether a user at a site will make a purchase at *any* future session. In a sense, this is a "macro" version of the session-level prediction problem.

To illustrate the difference in using site-centric data and user-centric data consider a user at a given point in time with three web sessions involving visiting Expedia. The prediction task involving site-centric data is to predict if a given user with n prior sessions with Expedia, will book at some future session based on the historical sessions $s_1, s_2, \ldots s_n$, where each s_i is of the form $Expedia_1$, $Expedia_2$, $Expedia_3$ for example. The prediction task involving user-centric data is to predict if a given user with n prior sessions with Expedia, will book at some future session based on the historical sessions $u_1, u_2, \ldots u_n$, where each u_i is of the form $Cheaptickets_1$, $Cheaptickets_2$, $Travelocity_1$, $Travelocity_2$, $Expedia_1$, $Expedia_2$, $Travelocity_3$, $Travelocity_4$, $Expedia_3$ for example. Hence for a user with N total sessions at Expedia, the site-centric and user-centric data would each contain N records, each with increasing historical content.

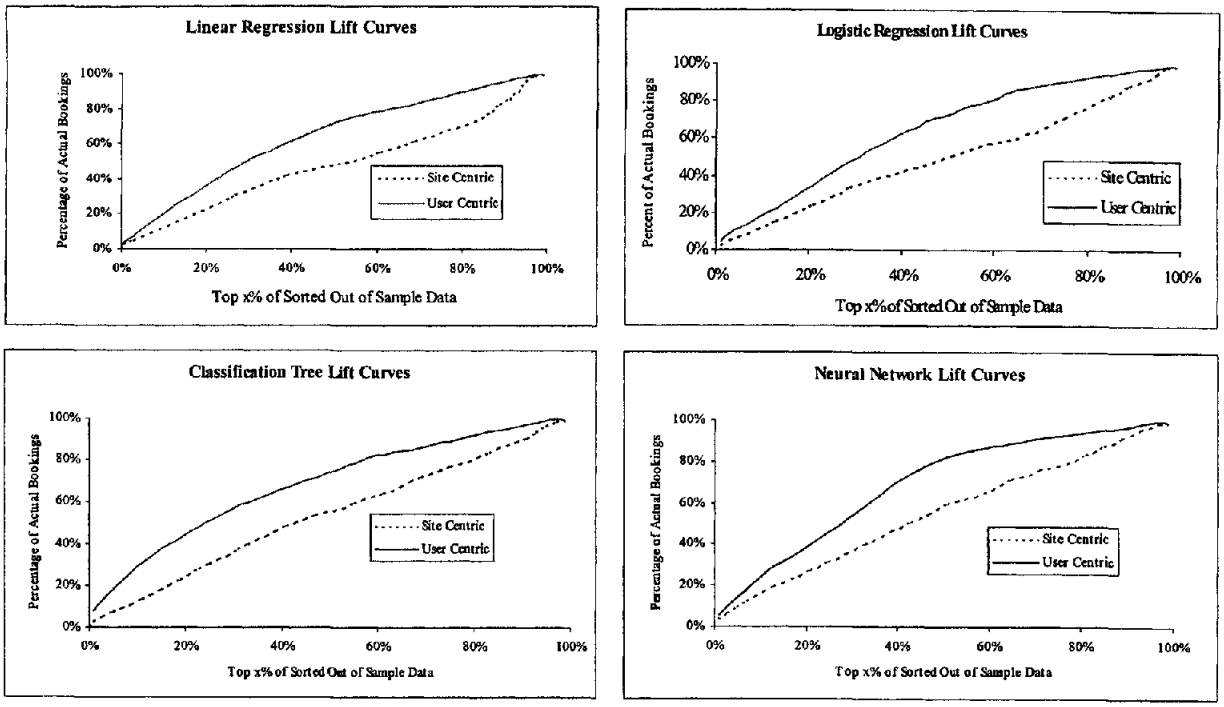
Unlike the previous problem, current session summaries are not explicitly created. These are subsumed in *summarize_historical*, since the prediction task is assumed to be at the end of any session. Hence the input to the user-level preprocessing algorithm are user sessions S_1, S_2,…, S_N and functions *summarize_historical* and *demographics* defined as follows:

1. summarize_historical(S, i), where S = {S_1, S_2,…, S_N}. This function is assumed to return summary variables based on all previous sessions and an indicator variable *booked* if the user booked in any (future) session from S_{i+1} to S_N. Note that the historical summaries are usually about the specific user in session S_i.
2. demographics(u_i) which returns the demographic information available about user u_i.

UserLevelDP, an algorithm for data preprocessing of site-centric and user-centric data for the user-level prediction task is presented in Figure 5.1. The preprocessing method creates a summary record at the end of each session based on increasingly additional historical session information. The specific metrics used in this problem are listed in the Appendix.

In this section we described the user-level prediction problem and presented algorithm UserLevelDP, for accurate preprocessing of clickstream data for user-level prediction tasks for site-centric and user-centric data. In the next section we present results from building various classifiers for both prediction tasks (session-level and user-level) and compare the results from doing so based on site-centric data to the results from user-centric data.

[3] Note that "user-level prediction" and "user-centric data" are very different. The terms "site-centric data" and "user-centric data" refer to the incomplete and complete versions of usage data. The terms "session-level prediction" and "user-level prediction" are used to describe 2 different problems that may be modeled from each of the 2 data sets.



Figures 6.1 - 6.4 (ordered clockwise from top-left). Lift Curve Comparisons – Session Level Prediction

6. RESULTS

Based on the 20,000 users' browsing behavior over the six month period, we selected 135 sites belonging to the five categories selling products mentioned in Section 3. From the user-level browsing data provided to us, for each of these sites, we constructed site-centric and user-centric datasets using the method described in Section 2. For each of the datasets we applied Probabilistic Clipping based methods with a desired data size of 2 Million (data for session-level prediction) and applied UserLevelDP (data for user-level prediction). Finally, based on all these preprocessed datasets, for each prediction problem we created two datasets: one aggregated site-centric preprocessed dataset and one aggregated user-centric preprocessed dataset. Each of these datasets are aggregate in the sense that they contains the union of records for all the sites under consideration. Below we summarize the results of building four different classifiers for the two problems. In each case, 40% of data was used in building classifiers, the remaining 60% was used as out of sample for performance comparison and evaluation.

6.1 Session-Level Prediction

The total number of records in the site-centric and user-centric datasets was 2 Million each and the number of explanatory variables was 16 for site-centric data and 41 for user-centric data. Figures 6.1 through 6.4 compare the lift curves of the different classifiers on the hold out samples for site-centric and user-centric data.

Note that for *all* the models, the lift obtained for models using user-centric data is significantly higher than the lift obtained from site-centric data. For example (see Fig. 6.4), the top 40% of the sorted out of sample data for the user-centric neural network classifier contains 70% of the actual booking sessions, while for the site-centric neural network model this contains only 47% of the actual booking sessions – a gain of 49%. A simple paired t-test shows that the lift generated from user-centric data, based on the lift at every two deciles, is significantly different from the lift from site-centric data (t_{15} = -14.03, P = 0.000). Similar results were obtained from numerous runs varying the training and hold out samples. That the same relative result holds for four very different model types (linear, loglinear and non-linear) and in experiments in which the learning sample (40%) was significantly smaller than the out of sample data (60%), provides initial evidence that the cost of using site-centric data alone in building models can be substantial.

There were several interesting qualitative findings based on analyzing each of the models derived, we present some representative examples here. For linear and logistic regressions, based on the site-centric model alone, the total time spent at a site in the past (*minutelh*) is significant and positively correlated with potential purchase. In the user-centric model though, *this effect is non-existent.* In the case of classification trees and neural networks, *minutelh*'s importance also drastically reduces in the user-centric case. Another contradiction is that site-centric approaches suggest that the total time spent at a current session (*minutelc*) is highly important – however, this effect disappears when user-centric data is used – are long sessions desirable or short sessions? Both these examples illustrate potential for erroneous conclusions based on models derived from site-centric data alone.

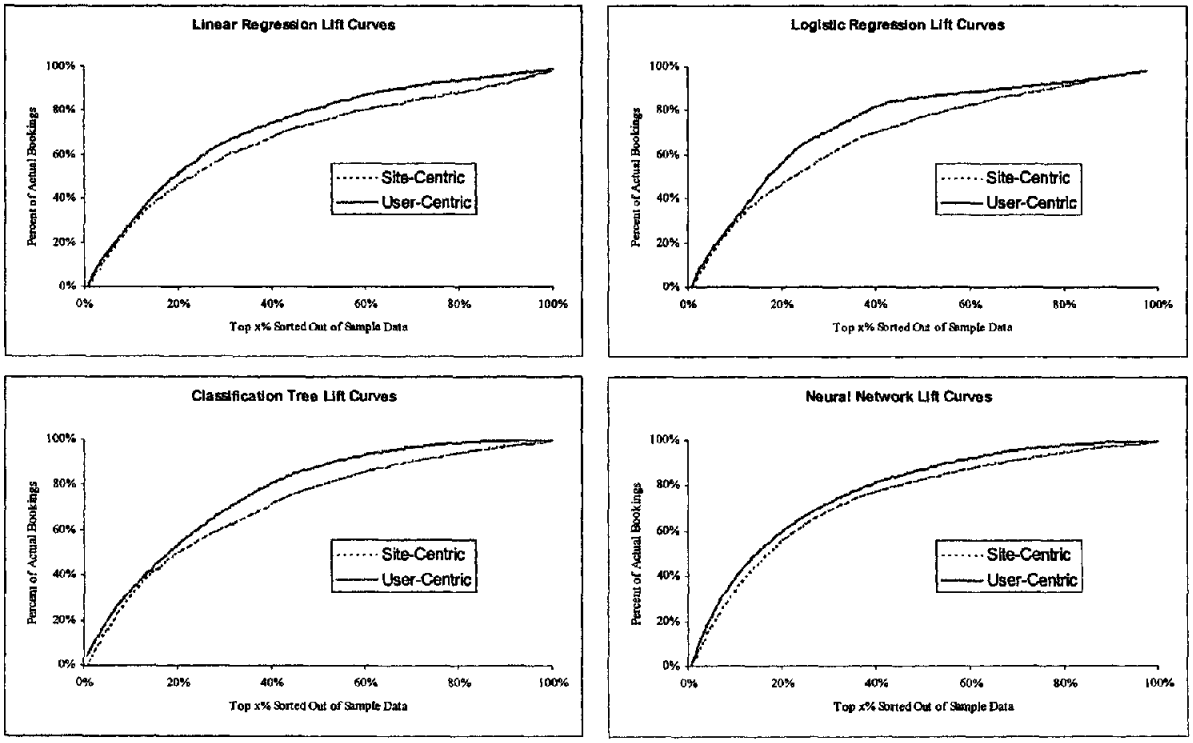
Further, the models from user-centric data indicate that in addition to the effects captured by site-centric models, there are several other factors, not captured in site-centric data, that have highly significant effects. First, bookings in the past at *any* site (*bookgh*) are very significant across all the models, and is the *most significant factor* in the neural net model. The market share of the user's past bookings captured by a site (*booksh*) is also highly significant across all the models and is the most significant factor in the logit model. Also the market share of the user's past sessions in which this site was the 'peak site' (*peakrate*) and whether in the current session a site is the 'peak site' (*path2*)) are both highly significant across all models. Finding such patterns can play a vital role in designing an effective online presence, but it is impossible to do so based on site-centric data alone.

6.2 User-Level Prediction

The total number of records in the site-centric and user-centric datasets was 801,367 each and the number of explanatory variables was 12 for site-centric data and 29 for user-centric data (see Appendix). Figures 6.5 presents overall predictive accuracies and booking class predictive accuracies on hold out data of site-centric and user-centric classifiers across 16 runs varying the records selected in the training and hold out samples but with the same 40/60 proportion (12% of the overall points were booking records). A paired t-test shows that accuracies for user-centric classifiers are significantly different.

Method	Run	Overall Pred. Accuracy		Booking Class Pred. Accuracy	
		s-centric	u-centric	s-centric	u-centric
Linear Regressions	1	88.2%	88.4%	5.30%	6.40%
	2	87.2%	87.6%	5.40%	7.30%
	3	87.4%	87.9%	5.40%	6.60%
	4	87.9%	88.3%	5.20%	6.90%
	5	88.2%	88.5%	5.50%	7.70%
Logit Models	6	88.40%	88.60%	11.70%	13.80%
	7	88.00%	88.30%	11.80%	14.70%
	8	88.20%	88.40%	12.20%	13.60%
	9	88.30%	88.60%	11.50%	13.90%
	10	88.60%	88.80%	12.00%	14.20%
Classification Trees	11	88.80%	89.50%	18.40%	23.00%
	12	88.60%	89.20%	16.20%	22.40%
	13	88.90%	89.70%	19.30%	24.50%
	14	88.60%	89.30%	17.80%	23.30%
	15	88.70%	89.30%	17.70%	23.70%
Neural Net	16	88.70%	89.90%	20.60%	29.30%
		t	**-7.10463**		**-6.5993**
		p	**5.96E-06**		**5.50E-06**

Figure 6.5. User-Level prediction accuracy comparison



Figures 6.6-6.9 (clockwise from top-left). User-Level prediction Lift-Curve Comparisons

Figures 6.6 through 6.9 compare the lift curves of the different classifiers on the hold out samples for site-centric and user-centric data. Based on a paired t-test comparing the lift at every two deciles, the user-centric classifiers generate significantly different lift than the site-centric classifiers (t_{15} = -8.81, P = 0.000). Again, for *all* models, the lift generated from user-centric classifiers significantly outperforms that generated from site-centric classifiers. However, as compared to the lift curves for session-level prediction, the differences are less pronounced. Direct comparison is hard since these are different problems. However, the results indicate that the effects of incomplete data may vary based on the personalization task considered.

7. CONCLUSION

The main results of the analyses are that models built from incomplete data (site-centric) are inferior to ones derived from complete data (user-centric). That this result holds for two different prediction problems and across four different classifiers that span the spectrum from linear to log-linear to non-linear provides initial evidence that these findings are robust. Further, qualitative analyses of the models provide evidence that potentially erroneous conclusions may be inferred from site-centric data and some findings may be ignored. Differences in the gains obtained for two different problems also suggest that the effects may vary based on the specific tasks considered.

In a sense, these results are not counter-intuitive – we would expect models derived from 'complete' data to be better than ones derived from incomplete data. Nevertheless, we argue that the results presented are significant for the following reasons:

1. Personalization models based solely on a site's logfile are the norm. Yet, little has been studied evaluating these models in the context of the incompleteness of data that they are learned on. Evidence presented in this paper suggests that in personalization from such data, what you don't know *can* hurt - Indeed, one of the key findings of this paper is that the magnitude of the difference between site-centric and user-centric approaches is strikingly high, which indicates that the cost of missing information may be much higher than is currently assumed.

2. Potentially erroneous conclusions can be formed on incomplete data. We present evidence that questions the blind application of such methods without explicit consideration of the impact of using incomplete information.

3. These results suggest that there may be value in business models that collect user-level data and provide these to individual sites to build personalization models. Some potential opportunities include customer opt-in models for licensing user-level data real-time to electronic commerce sites for building more effective personalization models.

In this paper, we empirically examined the implications of learning from incomplete web usage data in the context of two specific problems – session level prediction and user level prediction. For each of these problems we presented new algorithms for accurate data preprocessing of clickstream data and based on a comprehensive experiment on user-level clickstream data gathered from 20,000 users' browsing behavior, demonstrated that models built on user-centric data outperform models built on site-centric data for both prediction tasks.

8. ACKNOWLEDGMENTS

We would like to thank the Wharton e-Business Initiative (WeBI) and Jupiter Media Metrix for their support.

9. REFERENCES

[1] Adomavicius, G., and Tuzhilin, A., 1999, User Profiling in Personalization Applications through Rule Discovery and Validation, KDD-99, pp. 377-381, San Diego.

[2] Aggarwal, C.C., Sun, Z., and Yu, P.S., 1998, Online Generation of Profile Association Rules'. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining.

[3] Ansari, S., 2000. Integrating E-Commerce and Data Mining: Architecture and Challenges, Web-KDD, Aug., 2000.

[4] Brodley, C., and Kohavi, R., 2000, Peel the Onion, KDD-CUP 2000, Boston, 2000.

[5] Chan, P.K., 1999. A Non-Invasive Learning Approach to Building Web User Profiles. In Proceedings WebKDD 1999.

[6] Cutler, M, 2000, E-Metrics: Tomorrow's Business Metrics Today, In the Proceedings of the Sixth ACM SIGKDD International Conference on KDD, KDD 2000, Boston, Aug. 2000.

[7] Johnson, E., Moe, W., Fader, P., Bellman, S., and Lohse, J., 2000, On the Depth and Dynamics of Online Search Behavior, Wharton School Working Paper #00-014, June, 2000.

[8] Khabaza, T., 2001, "As E-asy as Falling Off a Web Log, Data mining Hits the Web", SPSS Data Mining Magazine, January.

[9] Kimbrough, S., Padmanabhan, B., and Zheng, Z., 2000, On Usage Metric for Determining Authoritative Sites, In the Proceedings of WITS 2000, Brisbane, Australia.

[10] Korgaonkar, P., and Wolin, L.D., 1999, A Multivariate analysis of Web usage, J. of Advertising Research, 39, pp 53-68.

[11] Mena, J., 1999, "Data Mining Your Website", Digital Press of Butterworth-Heinemann.

[12] Mobasher, B., Dai H., 2000, Discovery of Aggregate Usage Profiles for Web Personalization, Web-KDD, Aug., 2000

[13] Mobasher, B., Cooley, R., Srivastava J., 1999, Automatic Personalization Based on Web Usage Mining, Technical Report of Depaul University, TR 99-010.

[14] Moe, W., and Fader, P., 2000, Which Visits Lead to Purchases? Dynamic Conversion Behavior at e-Commerce Sites, The Wharton School, Working Paper #00-023. Aug. 2000 (A)

[15] Moe, W., and Fader, P., 2000, Capturing Evolving Visit Behavior in Clickstream Data, The Wharton School, Working Paper #00-003, Aug. 2000 (B).

[16] Novak, T., and Hoffman, D., 1997, New Metrics for New Media: Toward the Development of Web Measurement Standards, World Wide Web Journal 2(1), pp. 213-246.

[17] Nasraoui, O., Frigui, H., Joshi, A., Krishnapuram, R., 1999, Mining Web Access Logs Using Relational Competitive Fuzzy Clustering, In the Proceedings of the Eight International Fuzzy Systems Association World Congress, Taipei, August, 1999.

[18] Padmanabhan, B., Zheng, Z., Kimbrough, S., 2001, A Comparison of Site-Centric and User-Centric Data Mining Approaches to Predicting Session-Level Purchase Behavior on the Web, The Wharton School OPIM Dept Working Paper 01-01-03.

[19] Park, Y., Fader, P., 2000, Modeling Browsing Behavior at Multiple Sites, In the Proceedings of Informs Marketing Science Conference, Los Angels, June 2000.

[20] Perkowitz, M., Etzioni, O, 1997, Adaptive web sites: an AI challenge, In Proceedings of the 15th International Joint Conference on Artificial Intelligence.

[21] Perkowitz, M., Etzioni, O, 1997, Adaptive sites: Automatically synthesizing web pages, In Proc. of the Fifteenth National Conference on Artificial Intelligence (AAAI-98), pages 727--732, Madison, Wisconsin, July 1998.

[22] Pitkow, J., 1998, Summary of WWW Characterizations, Computer Networks And ISDN Systems(30:1-7), p551-558.

[23] Schechter, S., Krishnan, M., and Smith, M., 1998. Using Path Profiles to Predict HTTP Requests, In the Proceedings of the 7th Int'l. WWW Conference, Brisbane, Australia.

[24] Sen, S., Padmanabhan, B., Tuzhilin, A., White, N., and Stein, R., 1998. The Identification and Satisfaction Of Consumer Analysis-Driven Information Needs Of Marketers on The WWW, European Journal Of Marketing (32:7/8), pp. 688-702.

[25] Theusinger, C., Huber, K., 2000. Analyzing the Footsteps of Your Customers, Web-KDD 2000.

[26] VanderMeer, D., Dutta, K., Datta, A., 2000, Enabling Scalable Online Personalization on the Web, In the Proceedings of Electronic Commerce (EC00)/ ACM, Oct., 2000, Minneapolis.

APPENDIX: Variables for the User-Level Prediction Problem

A. Demographics

1 gender
2 age
3 income
4 education
5 household size
6 presence of children

B. Past History (Site-Centric)

7 No. of past bookings to this site so far
8 No. of sessions to this site so far
9 Time spent in this site so far in minutes
10 Average hits per session to this site
11 Average time spent per session to this site

C. Past History (User-Centric)

12 No. of past bookings at all sites so far
13 Average sessions per site so far
14 Total no. of sessions visited of all sites so far
15 Total minutes at all sites
16 Average hits per session
17 Average minutes per session
18 Total no. of unique shopping sites visited
19 Average no. of shopping sites visited per session
20 Percentage of single-site sessions
21 Percentage of total bookings to this site
22 Percentage of total hits to this site
23 Percentage of total sessions to this site
24 Percentage of total minutes to this site
25 No. of sessions starting with this site/total sessions of this site
26 No. of sessions the user spends the most time within this site/total sessions of this site
27 No. of sessions end with this site/total sessions of this site
28 No. of sessions start with search engines/total sessions of this site

D. Other Variables

29 Sub-category of the site (CD, books etc.)
30 Binary dependent variable indicating if the user has booked in the future