EVALUATION OF ONLINE PERSONALIZATION SYSTEMS: A SURVEY OF EVALUATION SCHEMES AND A KNOWLEDGE-BASED APPROACH

Yinghui Yang
Graduate School of Management
University of California, Davis
yiyang@ucdavis.edu

Balaji Padmanabhan
The Wharton School
University of Pennsylvania
mbalaji@wharton.upenn.edu

ABSTRACT

The use of personalization techniques for structuring online interactions with customers is common today. In a world in which a large part of customer interactions are done in this manner, systematic evaluation of these methods is critical. In this paper we study the problem of evaluating online personalization, and point out the difficulties of a scientific evaluation from automatically tracked data gathered by most online firms. A factor that contributes to the difficulty is that conducting true experiments may often not be possible due to the potential costs in doing so. When such true experimentation is not possible we present a systematic approach for evaluation that is based on bringing domain knowledge explicitly into the process. The advantage of our approach is that it presents a systematic approach to evaluate these systems by making explicit the domain knowledge. There is indeed no free lunch, and the disadvantage of the approach is that it relies on the accuracy of such domain knowledge. More generally this paper suggests that there may be problems in how online personalization systems are evaluated, and argues for systematic approaches to this important problem.

Keywords: Electronic Commerce; Personalization; Evaluation

1. Introduction

Businesses are investing considerable resources in developing and deploying personalization systems for customer interactions. Some cited reasons for doing so include a desire to develop stronger relationships with customers, improve profit margins and increase cross-selling [Schonberg et al. 2000, Cutler and Sterne 2000]. In the online world personalization methods are appealing for two key reasons: (1) their ability to differentiate users on a finer basis due to the capacity to collect more data and (2) their ability to do so in a scalable manner due to the automated nature of the techniques. According to a survey of 900 e-business executives conducted by Cahners In-Stat Group, large and medium-sized businesses nearly doubled their use of personalization technologies (for a list of personalization technologies, please see Pal and Rangaswamy, 2003) during 2003, with about 43% already having it in place by 2002. Global investment in personalization technologies is expected to reach $2.1 billion by 2006, up from $500 million in 2002, according to a recent report from Datamonitor.

There are several examples of personalization engines in use today. Amazon.com's personalization system, based on collaborative filtering, is well known. AT&T WorldNet[1] uses a method to decipher visitors' preferences unintrusively, and continually makes suggestions to visitors based on learned preferences. *DoubleClick* uses visitor profiles to target banner advertisements on their clients' sites that are more likely to be of interest to a specific visitor. *YesMail* specializes in targeting and sending personalized emails regarding special deals. Palm uses a recommendation engine that renders graphical product suggestions based on visitors' location on the site. In the business-to-business space, Dell Computer provides personalized Web pages for its corporate customers that simplifies placing and tracking orders.

As the above examples suggest, the use of personalization techniques for structuring online interactions with customers is common today. Three factors that enabled this are (1) the extremely large amount of user behavior data tracked at online sites, (2) the availability of powerful personalization techniques in commercial CRM systems, and

[1] See http://www.va-interactive.com/inbusiness/editorial/sales/ibt/personal.html

(3) the ease of implementing interaction strategies on the Web. These factors make it easy to build dynamic models (which can be implemented in real time) that automate interactions with customers. However, it is important to note that this ease of implementing personalized interactions online also raises the possibility of rolling out models with limitations that have not been understood and accounted for. Given the considerable amount of resources invested in building these systems, having a good method of knowing whether these systems are generating what is expected is important. Proper evaluation can also provide feedback to existing personalization systems and can be used to facilitate further enhancement of such systems. Hence, in a world in which many of the interactions with users are automatically structured, the systematic evaluation of such methods is critical.

This paper is a position paper on the importance of systematic evaluation of personalization methods. We point out how such methods are *currently* being evaluated and how such methods should *theoretically* be evaluated. A classical method of evaluation requires controlled experiments - a requirement that may be too expensive for some real applications. Instead, as we show in Section 3 current approaches use real data gathered in the natural setting to measure various metrics to evaluate the goodness of personalization systems. They often do not make a valid conclusion due to lack of a control group. In the absence of true experimentation, we present a knowledge-driven evaluation approach that has the characteristic of explicitly spelling out the knowledge assumptions and using real data gathered in natural settings. It is important to note that we *do not* advocate such an approach over classical experimentation, if such experimentation is at all possible. However in situations in which classical experimentation is infeasible or too expensive, the knowledge-driven approach can be a practical alternative, subject to the caveats discussed here. The advantage of our approach is that it presents a systematic approach to evaluate these systems by making explicit the domain knowledge. There is indeed no free lunch, and the disadvantage of the approach is that it relies on the accuracy of such domain knowledge. More generally this paper suggests that there are problems in how online personalization systems are evaluated and argues for systematic approaches to this important problem.

This paper is organized as follows. Section 2 describes the classical evaluation approach. Section 3 surveys existing approaches of evaluation. We present the knowledge-driven evaluation approach in Section 4, followed by a case study in Section 5. Section 6 concludes with a discussion.

2. The Classical Approach to Evaluation

First, in order to determine whether a personalization technique (p_1) is "good," some performance measure is necessary. This measure could be a metric such as customer satisfaction or number of products purchased, or the measure could be some other property that the system satisfies. For example, if the measure is customer satisfaction the personalization system will be "good" if it helps increase customer satisfaction. Then, in the context of evaluating a personalization technique p_1, two meaningful statements that can be made are:

1. p_1 is better than not personalizing (no p_1)
2. p_1 is better than some alternative personalization technique p_2.

In the first case the ideal experimental design [Campbell and Stanley 1966] would be:

R O_1 X_1 O_2
R O_3 O_4

In the second case the design would be.

R O_1 X_1 O_2
R O_5 X_2 O_6

R indicates random assignment of users to a group, each O_i indicates an observation that is made at a certain point in time, each X_i indicates the exposure of a group to some experimental condition, and each row represents a group of subjects. The left-to-right order is temporal, and the fact that any X_i or O_i is vertically above another X_i or O_i indicates that these exposures are done or measured simultaneously. Hence, the second design above indicates that there are two groups of randomly assigned users. At a specific point in time some observation, O_1 and O_5, is recorded for both groups. At a subsequent point in time both of these groups are exposed to different experimental conditions (exposure to personalization techniques p_1 and p_2 respectively). After this exposure, new observations, O_2 and O_6, are recorded for both groups.

In the first design, users are randomly allocated into a treatment group (with p_1) and a control group (no personalization), and in both groups a desired metric is observed before and after the treatment group intervention. For example, using this method to evaluate a firm's personalization technique would mean randomly choosing two groups of users, then observing number of purchases (or some other metrics) in the first period characterized by no personalization for both groups, followed by measuring purchases in the second period during which one of the

groups has personalized content. Comparing the measures in the two periods for both the groups would provide a measure for the effectiveness of the personalization system.

While this is the ideal experimental design, in practice such controlled experimentation may only be restrictively used in natural settings. The problem of carrying out such an approach in a natural setting is that most firms may be reluctant to have any reasonably long period with no personalization (or also a control group) or implementing an arbitrary second technique p_2, due to the risk of losing customers. In this scenario the data gathered automatically always reflects "$X\ O$" or "$O_1\ X\ O_2$"– i.e. always "personalizing", followed by data observation. Under no additional assumptions such data provides little information for scientifically evaluating personalization since it has no information on worlds without such personalization. In this situation (i.e. limited or no controlled experimentation), any method that attempts to evaluate personalization systems from data collected in a natural setting can provide a weak evaluation at best. Given the common nature of this problem, methods for weak evaluation are prevalent. A general characteristic of these approaches is that they cannot be used to prove that personalization is "good" due, notably, to the lack of an essential control group. In this sense, we refer to such methods as *weak evaluation schemes* and in the next section we discuss various examples of this.

3. Prior Work

In this section we review how personalization systems have been evaluated in the past and describe the limitations of common approaches to evaluation. We classify existing evaluation approaches along two dimensions. The first dimension is whether the personalization system is implemented in a real setting. While this is an important part of evaluation, few academic studies actually do so due to practical difficulties. The second dimension is how the measurement about the quality of the personalization systems is computed. While there are various ways to do this, for the purpose of the survey we group the various approaches based on whether the quality of the system is automatically computed from data generated (e.g., quality measures such as sales generated or click-through rate) or whether this is computed by using input from human subjects or experts (e.g., quality measures such as user satisfaction or consistency with user preferences). Table 1 presents a grouping of various approaches to evaluating personalization systems.

Table 1: A grouping of some of the approaches to evaluating personalization systems

	System implemented in a real situation	System not implemented in a real system
Quality of the system computed automatically from data	[Lawrence et al. 2001]	[Mobasher et al. 2000a, Mobasher et al. 2000b, Mobasher et al. 2001b]
Quality computed based on input from human subjects		[Geyer-Schulz and Hahsler 2002, Herlocker et al. 1999, Lin et al. 2002, Sarwar et al. 2000, Yu et al. 2001]

The first quadrant in Table 1 comprises evaluation approaches that implement a system in a real setting and evaluate the quality based on data collected as the system is used. Certain metrics measuring the performance of a personalization system can be measured without user interaction (e.g., sales generated, retention rate, click-though rate). Lawrence et al. [2001] describes a personalized recommender system (SmartPad) designed to suggest new products to supermarket shoppers based on their previous purchase behavior. The recommender functions in a remote shopping system in which supermarket customers use Personal Digital Assistants (PDAs) to compose and transmit their orders to the store, which assembles them for subsequent pickup. The recommender was subsequently used in a pilot program with several hundred customers at two Safeway stores. The authors use the fraction of orders containing at least one recommended product to evaluate the effectiveness of their personalized SmartPad. They also compared the revenue generated from the recommender to the revenue generated by products bought from the main "personal catalog" shopping list. This scheme of evaluation falls into the ($X_1\ O_2$) experimental setting when only the fraction of orders containing at least one recommended product is used. The second evaluation scheme involving comparison with a personal catalog has the structure:

$X_1 \quad O_2$
$X_2 \quad O_3$

For both cases, it is not possible to make a quality measure before the SmartPad is used. In the first case there is no control group for comparison and in the second case the evaluation approach compares the results from two

groups exposed to different conditions. Also unlike classical design, the subjects are not randomly assigned to groups but self-select into their groups.

More generally, studies proposing business performance-based metrics often refer to evaluation in this manner (i.e., evaluation is conducted in real situations and metrics are computed through automatically generated data). A substantial part of the literature on evaluating personalization focuses on identifying appropriate metrics (e.g., revenue, conversion rate, loyalty, click-through rate) that should be measured [Schonberg et al. 2000, Cutler and Sterne 2000, Lee et al. 2000a, Lee et al. 2000b]. None of these situations involves having a randomized control group.

The second quadrant comprises evaluation approaches that do not implement the system in a real setting but use historical data to generate quality measures [Mobasher et al. 2000a, Mobasher et al. 2000b, Mobasher et al. 2001b]. This is the most common approach reported in the data mining literature. Mobasher et al. [2000b] uses Web access logs to build an aggregate usage profile for personalization. Each transaction contains a list of pages viewed. For a given transaction in the evaluation set, Mobasher et al. [2000b] chooses a subset of pages viewed, produces a recommendation set, and compares this recommended set with what the user actually did in the remainder of the transaction. Typically, the system is considered good if its recommendations are consistent with the user's actions that were captured in a scenario in which the personalization system was not used. In general, this is a problem, since the user's behavior may indeed be affected by the system.

The third quadrant comprises evaluation approaches that implement a system in a real setting and evaluate the quality based on input from users. Certain quality metrics (e.g., ease of use and user satisfaction) require interaction with the user. This type of evaluation is often conducted via a survey of system users. This is an approach used by some online stores that conduct a user satisfaction survey at the end of a user's visit. Many times, this survey is conducted by an external organization, such as Bizrate. Common approaches that fall into this quadrant have the problem that there is typically no control group. While several approaches in the real world fall into this category, few academic papers use such an evaluation approach.

The fourth quadrant comprises evaluation approaches that do not implement the system in a real setting and evaluate the quality based on input from users. In this category, recommendations are typically made based on historical usage data. Evaluation is conducted by collecting input from sampled human subjects and comparing the human input with the recommendation list. Geyer-Schulz and Hahsler [2002] used historical usage data containing a list of itemsets capturing products bought together. For evaluation, Geyer-Schulz and Hahsler [2002] produced for each item a list of other items that co-occurred with the first item in at least one transaction. Geyer-Schulz and Hahsler [2002] randomly selected 300 lists, asked students and researchers whether they would recommend the later items to a person interested in the first one, and then compared that with the results of the recommender algorithms. They also surveyed some commonly used performance measures for recommendation algorithms, such as precision, recall, F-measure, accuracy, coverage, and mean absolute error. Herlocker et al. [1999] used the historical movie rating data from the MovieLens movie recommendation site. The data consisted of ratings from 1,173 users, with every user having at least 20 ratings. Ten percent of the users were randomly selected to be test users. From each user in the test set, ratings for 5 items were withheld and predictions were computed for those 5 items using their algorithms. The prediction results were compared with the true ratings. Studies in Lin et al. [2002], Sarwar et al. [2000] and Yu et al. [2001] used the same methodology on movie rating data. There are three main drawbacks to such approaches: (1) The system is not implemented in a real setting, (2) data on ratings of various alternatives from a large group of users is typically not available, and (3) the historical data is collected from users in a world in which the personalization system was not used (a problem shared by approaches in the second quadrant).

We note that recent work has also made a similar point about the need to have better designs when evaluating personalization systems. Specifically, both Spiliopoulou [2000] and Lawrence et al. [2001] pointed out the difficulty in the evaluation of personalization systems. "*In order to quantify the impact of the recommender system, it would have been useful to have a control group of customers who received "placebo" recommendations, such as a list of randomly chosen products. This approach was not feasible, however, since we were dealing with a live system with real customers doing real shopping.*" [Lawrence et al. 2001] "*In principle, we have an evaluation methodology. But it comes with a high price. We must gather a group of test persons, ensure they are representative of the target group of customers and establish an experimental environment for them to work. Although some companies can afford the time and cost of this process when they establish a site or perform a fundamental redesign, very few can launch the same process each time they want to monitor the site's quality. . . . Ideally, we would like to evaluate a site based on the data automatically recorded on it.*" [Spiliopoulou 2000]

To summarize, most of the approaches used in the literature to evaluate online personalization systems do not use a classical experimental design. Hence these evaluation approaches cannot be used to prove that personalization is good due, notably, to the lack of a control group. In this sense, we refer to such methods as *weak evaluation*

schemes. In the next section we suggest that good prior knowledge can offset some of the limitations that arise in evaluation and present a systematic method that can be used formally to evaluate personalization under a strict assumption that the knowledge is correct, consistent, and adequate.

4. Knowledge-Driven Evaluation

The motivation behind our approach is that good prior knowledge can be used to partially offset limitations of the data (that arise from implicitly using an "*X O*" design only) in the context of evaluating personalization. The issue is: what should this knowledge be and how can it be used in conjunction with the data to evaluate personalization? Hypothetically, if a true experiment were conducted with a treatment group (personalization) and a control group (no personalization), observed differences could be used to evaluate the effect of personalization. Hence, it stands to reason that knowledge about *expected outcomes* under various personalization scenarios could be used to evaluate personalization. For example, consider the following knowledge about expected outcomes under different scenarios:

1. If the personalization is good, then the product ranking will be a significant positive predictor of purchasing the product.
2. If the personalization is bad, then the product ranking will be a significant negative predictor of purchasing the product.
3. If the personalization is irrelevant, then the product ranking will not be significant in predicting purchase.

Note that the connection with using real data that is tracked at a site using a specific personalization technique is that the consequents (RHS) of these rules are statements that can be evaluated on the data. The simple case above is useful for two related reasons. First, it provides some intuitions into what knowledge is necessary for evaluating personalization (knowledge about expected outcomes). Second, it provides a manner of determining how this knowledge should be used. In the example, the set of rules indicates that measuring the significance of product rankings provides adequate discriminatory power to distinguish which scenario holds. In this paper we term such conditions as *distinguishing sets* for a scenario.

Generalizing from this example, our knowledge-based evaluation approach consists of three stages: (1) generating knowledge about various scenarios, (2) determining what to measure in order to obtain sufficient discriminatory power, (3) making this measurement to determine which scenario holds. Below we formalize these ideas.

Assume that $s_1, s_2 \ldots, s_p$ are mutually exclusive and collectively exhaustive scenarios. For example, in the case of evaluating a personalization technique, the number of scenarios is three: personalization is good, bad, or irrelevant. The knowledge consists of rules that represent expected outcomes in the various scenarios. Each rule is in the form: if s_i then a_j, where s_i is a scenario and a_j is a single atomic condition based on the set of literals in the domain.

Definition 1. A set of statements $d = \{ a_1, a_2 .., a_k\}$ is a *distinguishing set* for a scenario s_i if $d \rightarrow \neg s_j$ for all $s_j \neq s_i$ and it is not the case that $d \rightarrow \neg s_i$.

Intuitively, a distinguishing set for a scenario is one that rules out all other scenarios except the target scenario. Note that the rules are in the form $s_i \rightarrow a_j$. For each such rule, it logically follows that $\neg a_j \rightarrow \neg s_i$ (these rules therefore can be used to determine conditions under which any set of scenarios can be shown not to hold). In general there may exist several such sets for a scenario, and in order to make the measurement process parsimonious we introduce the notion of a *minimal distinguishing set*.

Definition 2. A set of statements $d = \{ a_1, a_2 .., a_k \}$ is a *minimal distinguishing set* for a scenario s_i if d is a distinguishing set and there exists no $d' \subset d$ such that d' is also a distinguishing set for s_i.

Example 1. Consider the following sets of rules, where p is a personalization technique. The scenarios of p are good, bad, or irrelevant and the consequents A, B, and C are assumed to be binary statements that can be evaluated based on gathered data (Table 2).

Table 2: An example for acquired knowledge about personalization

If p is good then A	If p is bad then A	If p is irrelevant then ¬A
If p is good then B	If p is bad then ¬B	If p is irrelevant then B
If p is good then C	If p is bad then C	If p is irrelevant then C

In this example, no single condition can uniquely establish p as being good, distinguished from the other possibilities. {A, B} and {A, B, C} are distinguishing sets for p being good. Although {A, B} is the only minimal

distinguishing set for the scenario in this example, in general there can be several minimal distinguishing sets for a scenario.

Definition 3. A distinguishing set *d* holds for a scenario *s* if all the statements in *d* can be evaluated to be true on gathered data.

Based on these concepts, a scenario (e.g., "personalization is good") can be evaluated by determining the set of minimal distinguishing sets and testing whether any distinguishing set holds. Determining the set of minimal distinguishing sets for a target scenario is an interesting problem and one that is important as the number of rules describing the scenarios increases.

Figure 1 presents the general algorithm that determines the set of minimal distinguishing sets for a scenario. The strategy of the algorithm is to use two ideas that help to focus the search on only the minimal sets:

(1) Iteratively generate the minimal distinguishing sets starting from the smallest sets. In any iteration, if a set is determined to be a distinguishing set, it is added to the minimal set and immediately deleted from the consideration set, since supersets of this set cannot be minimal. The process is similar to the itemset generation procedure used in rule discovery techniques [Agrawal et al. 1995].

(2) Note that a minimal distinguishing set cannot contain $a_{k,}$ if a_k does not rule out at least one of the other scenarios. Hence, the first iteration must start with only the set of all single atomic conditions that contradict at least one condition in the RHS of some rule in the other scenarios.

Input: Domain rules *R*, scenarios *S, target scenario s*
Output: set of minimal distinguishing sets *DS*
DS = {}, k=0
C_k = { m | m is a set consisting of a single atomic condition which contradicts some *a* where $s_1 \rightarrow a$ and $s_1 \in S\text{-}s$, and does not contradict any *b* where $s \rightarrow b$}
Repeat {
  for each $c \in C_k$ {
    if *c* is a distinguishing set then {
      DS = DS $\cup$ {*c*}
      $C_k = C_k - \{c\}$
    }
  }
  C_{k+1} = { $c \mid c = c_1 \cup c_2$ where $c_1, c_2 \in C_k$, and c_1 and c_2 do not contradict with each other}
  k = k + 1
} while (C_k different from C_{k-1})
Output 'DS'

Figure 1: Generating minimal distinguishing sets

If the rule is of the form "if s_i then (a_j and a_k)", it can certainly be separated into two rules with atomic conditions: (if s_i then a_j) and (if s_i then a_k). The same applies to rules with more than two conditions. If the rule is of the form "if s_i then (a_j or a_k)", it logically follows that ($\neg a_j$ and $\neg a_k \rightarrow \neg s_i$). When calculating the minimal distinguishing set for scenarios other than s_i, C_0 can contain a composite condition ($\neg a_j$ and $\neg a_k$) if it does not contradict any *b* where $s \rightarrow b$ (*s* is the target scenario).

In this section we presented an approach to generating minimal distinguishing sets. A few important characteristics of this approach should be emphasized:

(1) This approach treats the evaluation problem in a more systematic manner by making the assumptions explicit. We believe that this is important, given the natural inadequacies of the data in addressing the problem of evaluation.

(2) The conditions that form the knowledge do not have to be "metrics" in the sense usually considered in prior work. They can be other indicators, such as "the rankings of the system significantly affect the likelihood of purchase." These indicators could capture problems even in situations in which baselines are not available to determine the goodness of conventional metrics.

(3) The approach allows the use of several possible rules that may generate multiple distinguishing sets. This can permit evaluation under more flexible circumstances.

(4) Existing approaches, such as the methods listed in prior work at the end of Section 3, that essentially look at metrics *can* be acceptable methods of evaluation if certain assumptions hold. For

example, under the knowledge that *if s=good then clickthroughrate=higher*, *if s=bad then clickthroughrate=lower*, *if s=irrelevant then clickthroughrate=same*, looking at trends in *clickthroughrate* is (trivially) an acceptable method of evaluation under our approach.

(5) Rules embedded in personalization systems are common. In addition, our approach advocates using a combination of automatically tracked data and explicit knowledge (in the form of evaluation rules) to continuously evaluate personalization techniques.

Three key assumptions are made about the quality of the domain knowledge:

(1) It is consistent. That is, for any realization that is possible in the real world, it is not possible to conclude *x* and (not *x*) for any *x*.

(2) It is adequate. That is, it results in at least one distinguishing set if any exists indeed.

(3) It is correct.

These are strict assumptions that must be made, given the fact that an evaluation has to be made without conducting a true experiment. Measuring the quality of knowledge to determine whether these assumptions are satisfied is a hard problem.

In this section we presented a new approach to evaluating personalization that uses domain knowledge to offset some of the limitations regarding the inadequacy of the data. In the next section we present a brief summary of results studying the use of this approach in evaluating personalization at an online firm selling telecommunications services to subscribers.

5. Case Study

In this section we present a simple example of using minimal distinguishing sets within a knowledge-driven evaluation framework. Although the case is about a real firm with a personalization system, the knowledge representing the expected outcomes here is quite simple and, for this particular case, there is one trivial minimal distinguishing set. However, the case is useful in that it illustrates how the method can be used in conjunction with real data.

We use data collected at an online firm that provides a decision platform that helps customers select products or services. When customers enter the site they are asked to provide some information about themselves. The system generates a ranked list of recommended products (such as wireless plans, for example) according to customer information such as profiles and preferences, and product information, such as pricing, performance, and features that the company stores in its database. The plans that the customer is more likely to buy are ranked higher on the recommended list.

The variables are partitioned into customer variables (all of the survey variables to which the customers respond), product variables (all of the variables that capture features of the products, such as number of roaming minutes in a certain plan for example), the rank of each product in the returned list of recommendations, and a binary variable representing whether each product was added to a shopping cart. The data consisted of 1.7 million records, each with 34 attributes representing a customer's choice regarding a certain recommended product. It contains information about the customer (26 attributes), the product (6 attributes), the ranking of the product (1 attribute), and whether the customer added the product to a shopping cart (1 attribute).

In this case there were no controlled experiments done and there was only data gathered in the natural setting – a common occurrence in practice. Hence given that the system could not be evaluated using controlled experiments a knowledge-based approach was considered. As described in Section 4, the idea is to use assumptions / prior knowledge about expected outcomes to evaluate this system. In this case the initial knowledge about expected outcomes is simple, and is listed in Table 3.

While in Section 4 we presented a general approach to determining the distinguishing sets, in this case study the distinguishing set can be identified trivially due to the small set of rules comprising the prior knowledge or assumptions used. First note that the knowledge indicates that product and customer information will *always* be significant predictors of purchase - regardless of whether the personalization is good - since it indicates only that customers buy products that match their interests (this can be seen from reading the first two rows of the table). Hence in our terminology, none of these attributes are part of a "minimal distinguishing set" since they do not help distinguish between when personalization is "good", "bad" or "irrelevant". In other words, even if the company was using a badly designed personalization system the basic customer attributes will still be expected to be predictors of purchase.

Table 3: Initial Knowledge

If *p* is good, then customer attributes will be relevant in predicting purchase	If *p* is bad, then customer attributes will be relevant in predicting purchase	If *p* is irrelevant, then customer attributes will be relevant in predicting purchase
If *p* is good, then product attributes will be relevant in predicting purchase	If *p* is bad, then product attributes will be relevant in predicting purchase	If *p* is irrelevant, then product attributes will be relevant in predicting purchase
If *p* is good, then rankings will be significant in a positive manner in predicting purchase	If *p* is bad, then rankings will be significant in a negative manner in predicting purchase	If *p* is irrelevant, then rankings will be insignificant in predicting purchase

The only distinguishing set is whether *rankings* are significant (as can be seen by reading the third row of the table). The three related rules that are part of the prior knowledge (or assumptions) state that in a good personalization system customer purchase is *positively* correlated with the ranking of the product, whereas in a bad system it will be *negatively* correlated, and in a random system it will be insignificant. While our approach just assumes this knowledge as input, the rules listed here are intuitive in that they capture the notion that when a higher ranked product is purchased (and therefore deemed "good") by a customer the system can be considered an effective one.

Once a distinguishing set is identified then measuring its value is all that needs to be done to determine if the personalization system is good. Next note that the consequents of all the rules in the knowledge base are conditions that can be tested by building models from the data gathered in the natural setting. For instance determining whether the rankings are indeed significant in predicting purchase can be done by building models from the real data gathered in the natural setting. Note that *how* the condition (i.e. distinguishing set) is tested is not the main purpose of this paper. While any appropriate model can be built, here we show how a non-parametric model (classification trees) can be used for this purpose.

One technique for determining the value of the distinguishing set (i.e. how rankings relate to purchase) is to build classification trees to predict purchase with and without rankings. Looking at the models and the difference in the lift curves (a common performance measure used when building classification trees on highly skewed binary data) on out-of-sample data in the two cases can provide a measure of how significant the rankings are.

Three predictive models were built.

Model 1: Shopping cart addition is modeled as a function of customer-specific variables alone. This is a useful prior that can provide the likelihood of purchase based on input preference alone.

Model 2: Shopping cart addition is modeled as a function of customer-specific variables and product-specific variables. The improvement of Model 2 over Model 1 can provide a measure of how well the personalization system selects a set of plans from the space of all possible plans.

Model 3: Shopping cart addition is modeled as a function of customer-specific variables, product-specific variables, and the ranking of the product. The improvement of Model 3 over Model 2 can provide a measure of the usefulness of the ranking returned by the personalization system.

Classification trees iteratively split the data based on a single variable, and in general the variables closer to the root of the tree are important predictors of the outcome. The trees drawn in Figure 2 and Figure 3 are the first two levels of the decision trees corresponding to the model 1 and model 3. The first split of the tree for model 1 is based on variable A1 (we do not provide the exact variables for confidentiality reasons, but what is important in this case study is how the "ranking" variable makes its way into the tree in the next case), and the first split of the tree for model 3 is based on the "ranking" variable. In Figure 4, the left-most lift curve corresponds to model 3, the middle lift curve corresponds to model 2, and the right-most curve corresponds to a random model. In general the higher the lift curve is over the random lift curve, the better the model is. There was a 5-10% improvement in lift of Model 3 over Model 1 indicating that ranking is significant in predicting purchase, and the decision tree showed that the effect was positive. The decision trees have interesting qualitative results that show how certain variables influence the likelihood of purchase. The tree corresponding to Model 3 splits first on ranking. The subset with ranking=1 has a prior of 80.6% (a gain of 30.6% over the default percentage because the data was scaled to a 50-50 prior for the dependent variable) indicating the ranking is indeed relevant to the dependant variable.

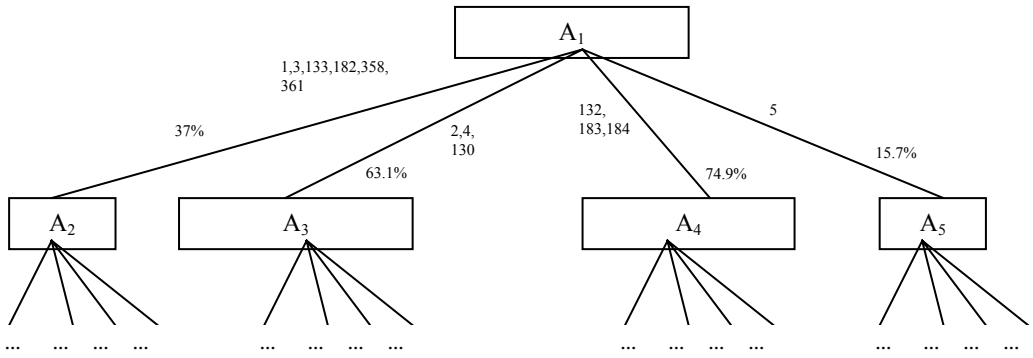

Figure 2: The first level of the decision tree for Model 1

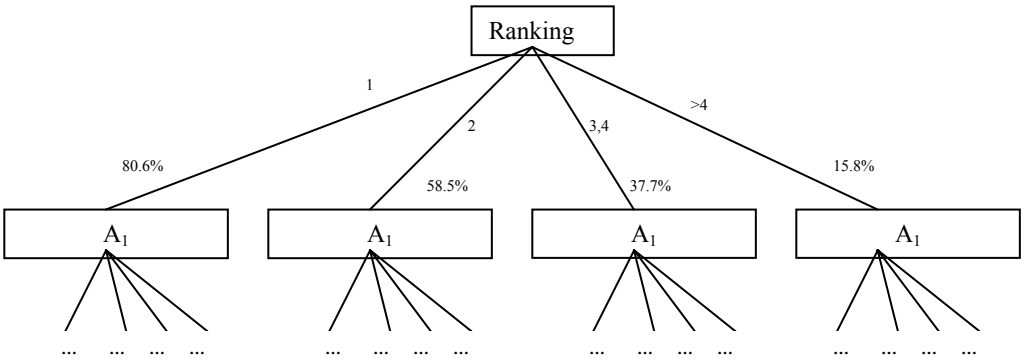

Figure 3: The first level of the decision tree for Model 3

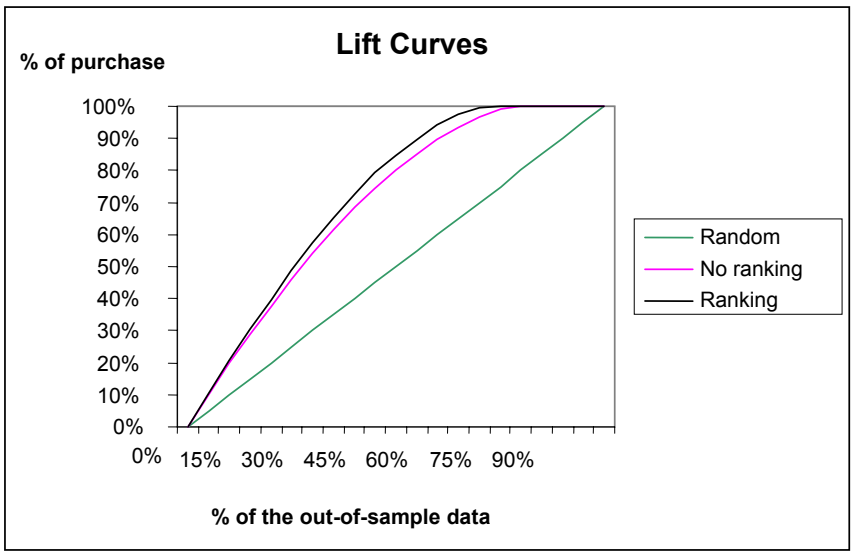

Figure 4: The lift curves of Models 2 and 3

To summarize, the main observation is that the first split in the classification tree built for Model 3 was based on *rankings*, where the higher rankings corresponded to greater purchases. The lift curves on 50% out-of-sample data indicated that the lift was higher than the case without rankings, and the average increase in lift was

approximately 5%. Hence rankings were measured to be a significant and positive predictor of purchase. Given that the distinguishing set is now determined (and measured), based on table 3 it can be concluded that the personalization system is "good", subject to all the caveats described in this paper.

This example demonstrates the advantage of using knowledge (and making it explicit) in evaluating personalization. Indeed, it is important to note that there is no free lunch here: The quality of the initial knowledge essentially determines how "correct" the evaluation is. This is unavoidable, given the inherent difficulties in making statements about evaluation from data.

6. Discussion

In this paper we studied the problem of evaluating online personalization and pointed out the difficulties of a scientific evaluation from automatically tracked data gathered by online firms. Based on reviewing the literature, we show that most conventional approaches for personalization do not conduct real experiments for validation. A significant factor that affects this is the difficulty in conducting true experiments due to the potential costs in doing so. Under this circumstance we point out that existing methods for evaluation are inherently weak. To offset the limitations we use domain knowledge to evaluate personalization. In particular, we present a systematic approach for using prior knowledge for the evaluation problem and discuss the advantages of this approach.

Although much of this paper describes a knowledge-driven approach that can be used on automatically tracked data, as mentioned in the introduction, we do not advocate this approach over classical experimentation, if such experimentation is at all feasible. From a methodological point of view, conducting controlled experiments is the ideal approach to determining whether a personalization system is good. However, from a practical perspective, conducting real experiments often enough may be difficult, and it is appealing to use continuously collected data to evaluate personalization approaches. In this paper we present a scenario under which this data *can* be used, but interpretation of results must be done strictly in the context of the knowledge used.

A clear advantage of evaluation based on controlled experiments is its generality in being able to evaluate a wide range of systems. If the goal is properly designed into the experiment, and the customers chosen in the survey are representative, it is often an ideal method for evaluating personalization systems. Sometimes personalization systems are designed to enhance customer satisfaction, and directly asking customers about how they feel is the best way of knowing whether the goal is achieved. There are also cases where data automatically generated is not enough to properly evaluate the personalization system. Knowledge-based evaluation approach can provide value when automatically generated data and enough knowledge are available. When the knowledge is not enough or accurate, such an approach should not be adopted. However this approach is easy to use once enough data and knowledge are in place. Given this advantage, it can be especially valuable for companies that need to continuously monitor their system via automatically generated data. From a practical perspective, the knowledge may just be viewed as assumptions, and the approach presented here can be viewed as one that requires making assumptions explicit when using data gathered in a natural setting to evaluate systems.

In a broader sense, the issue of evaluating personalization is significant, and much more research is necessary to address this issue. Particularly some of the relevant issues are methods that evaluate the quality of the knowledge, alternate methods for using prior knowledge in the evaluation process, and studying the evaluation under dynamic scenarios in which knowledge and personalization techniques evolve over time. Another key issue that needs to be considered during evaluation is the manner in which the personalization system is implemented – whether, for example, it is based on modifying content shown to a user or based on providing different navigation paths for a user. This is an issue that is not considered explicitly in this paper, but clearly the two approaches may be evaluated using very different approaches. Hence another interesting topic for future work is to examine the different methods in which personalization is provided to provide guidelines for evaluation in each case.

REFERENCES

Agrawal, R., H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast Discovery of Association Rules," Advances in Knowledge Discovery and Data Mining, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (eds.), MIT Press, Cambridge, MA, pp. 307–328, 1995.

Campbell, D. and J. Stanley, Experimental and Quasi-experimental Designs for Research, Rand McNally college publishing company, Chicago, IL, 1966.

Cutler, M. and J. Sterne, "E-Metrics - business metrics for the new economy," Net Genesis Corp., http://www.targeting.com/emetrics.pdf, 2000.

Geyer-Schulz and A., M. Hahsler, "Evaluation of Recommender Algorithms for an Internet Information Broker based on Simple Association Rules and on the Repeat-Buying Theory," Proceedings of the. 4th WebKDD

Workshop: Web Mining for Usage Patterns & User Profiles, pp. 100-114, Edmonton, Alberta, Canada, July 2002.

Herlocker, J., J. Konstan, A. Borchers, and J. Riedl, "An algorithmic framework for performing collaborative filtering," Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 230-237, Berkeley, CA, August 1999.

Lawrence, R. D., G. S. Almasi, V. Kotlyar, M. S. Viveros, and S. Duri. "Personalization of supermarket product recommendations," *The International Journal of Data Mining and Knowledge Discovery (Special Issue on Applications of Data Mining to Electronic Commerce),* Vol. 5, pp. 11-32, 2001.

Lee, J., M. Podlaseck, E. Schonberg, R. Hoch, and S. Gomory, "Analysis and visualization of metrics for online merchandising," Lecture Notes in Computer Science: Advances in Web Usage Tracking and Profiling, Springer-Verlag, New York, NY, 2000a.

Lee, J., M. Podlaseck, E. Schonberg, R. Hoch, and S. Gomory, "Understanding merchandising effectiveness of online stores," *The International Journal of Electronic Commerce and Business Media*, Vol. 10, No. 1, pp. 20-28, 2000b.

Lin, W., S. A. Alvarez, and C. Ruiz, "Efficient adaptive-support association rule mining for recommender systems," *Data Mining and Knowledge Discovery*, Vol. 6, pp. 83-105, 2002.

Mobasher, B., B. Berendt, and M. Spiliopoulou, "KDD for Personalization," Tutorial at the 12th European Conference on Machine Learning (ECML01) / 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD01), Freiburg, Germany, September 2001a.

Mobasher, B., R. Cooley, and J. Srivastava, "Automatic Personalization Based On Web Usage Mining," *Communication of ACM,* Vol. 43, No. 8, pp. 142-151, 2000a.

Mobasher, B., H. Dai, T. Luo, and M. Nakagawa, "Improving the Effectiveness of Collaborative Filtering on Anonymous Web Usage Data," Proceedings of the IJCAI 2001 Workshop on Intelligent Techniques for Web Personalization (ITWP), Seattle, Washington, August 2001b.

Mobasher, B., H. Dai, T. Luo, Y. Sung, M. Nakagawa, and J. Wiltshire, "Discovery of Aggregate Usage Profiles for Web Personalization," Proceedings of Web Mining for E-Commerce Workshop (WebKDD), Boston, MA, August 2000b.

Pal, Nirmal and Arvind Rangaswamy, The Power of One: Gaining Business Value from Personalization Technologies, 2003.

Sarwar, B., G. Karypis, J. Konstan, and J. Riedl, "Application of dimensionality reduction in recommender systems– a case study," Proceedings of Web Mining for E-Commerce Workshop (WebKDD), Boston, MA, August 2000.

Schonberg, E., T. Cofino, R. Hoch, M. Podlaseck, and S. L. Spraragen. "Measuring success," *Communications of the ACM,* Vol. 43, No. 8, pp. 53-57, 2000.

Spiliopoulou M. "Web usage mining for site evaluation: Making a site better fit its users," *Communications of ACM* , Vol. 43, No. 8, pp. 127-134, 2000.

Yu, K., X. Xu, M. Ester, and H. Kriegel, "Selecting relevant instances for efficient accurate collaborative filtering," Proceedings of the tenth International Conference on Information and Knowledge Management (CIKM), pp. 239-246, Atlanta, Georgia, November 2001.