Personalization using Hybrid Data Mining Approaches in E-business Applications

Olena Parkhomenko, Chintan Patel, Yugyung Lee

School of Computing and Engineering
University of Missouri – Kansas City
{ophwf, copdk4, leeyu}@umkc.edu

Abstract

Effective personalization is greatly demanded in highly heterogeneous and diverse e-commerce domain. In our approach we rely on the idea that an effective personalization technique has to be customized to meet the specific needs of every particular domain and deliver quality recommendations. With this in mind, we have combined Bayesian classification methods with association rule mining to model individual customer's behavior. While Bayesian classifier is for effective customer profiles, rules-based analysis works for both customer and non-customer objectives, such as reducing over-stocked items. This paper also presents a comparative analysis of the existing personalization techniques for the improvement of a distributed online customer care application. In this paper, we have successfully demonstrated on the example of the SprintPCS customer care domain that our approach is an efficient recommendation model for the online customer care.

Introduction

Today, customers have more choices than ever. They are more aware of the possibilities and more demanding of personal attention. This situation shifts the focus from the product toward the individual customer. The more personal this becomes, the more customers will be loyal to an organization. It is now becoming increasingly important for a company to build a strong relationship with its customers.

This is where personalization technology steps in. It can allow a company to customize content, sales offers and loyalty programs each time a customer came to the site. Customers can be classified into "buckets" based on past behavior as well as predictions of future behavior. Each of these "buckets" can then be treated differently, based on marketing needs and behavior, providing the individual attention and offers that match customers' real interests. However, this powerful scenario still suffers from pitfalls of implementation, since most of the existing personalization engines have shortcomings and limitations and generally speaking are not able to provide the expected performance level. The reason for this is that the existing personalization methods fail to provide a universal solution that could satisfy the needs of any problem domain. As a result, most of today's personalization software is custom-build, which consumes a lot of effort and resources, and in fact, could be compared to re-inventing the bicycle.

With this in mind, we have proposed the following personalization solution strategy: On the first step we identify a family of e-business applications that share similar needs and requirements for personalization. In other words, we can identify a particular problem domain. On the second step we create a model, based on the previous research, comparative analysis and evaluation metrics, and formulate a concrete implementation approach for the problem domain.

In this paper we have conducted a comparative analysis of the existing personalization techniques while focusing on the strategy that can be successfully applied to improve a particular problem domain - distributed online customer care. Though there exist various personalization techniques, such as collaborative filtering, rule-based analysis and data-mining methods that are currently used in e-business applications, there are still drawbacks and issues to be solved, such as generating effective customer profiles and providing accurate recommendations. In our approach we attempt to prove the idea that an effective personalization technique has to be customized to meet the specific needs of every particular domain, and then only it would be able to deliver quality recommendations and thus serve its purpose. With this in mind, we have combined Bayesian classification methods with association rule mining to model individual customer's behavior. While Bayesian classifier can be successfully used to create effective customer profiles, rules-based analysis would allow action targeted for both customer and non-customer objectives, such as reducing over-stocked items. Based on the previous research, we consider that these two strategies complement each other and add up to an efficient recommendation engine that is capable of producing accurate recommendations for the online customer care problem domain.

Related Work

The most simple and straightforward personalization approach is clickstream analysis, where the major issue is the problem of sequences of clicks (Andersen et al. 2000).

Next comes probably the most popular personalization technology of today's market - collaborative filtering. Pure collaborative filtering suffers from a variety of limitations, such as scalability and effectiveness in the face of very large and sparse data sets. Offline clustering of user transactions can significantly improve the efficiency of such systems, however, at the cost of decreased accuracy. In the case of anonymous web usage data there is also the challenge of accurately predicting user interests based on very short user clickstream trails, and without the benefit of more detailed user information (Mobasher et al. 2001). Clustering and data-partitioning algorithms in collaborative filtering can potentially improve the quality of collaborative filtering predictions and increase the scalability of collaborative filtering systems (O'Connor and Herlocker 1999).

GroupLens (Sarwar et al. 1998) implemented a hybrid collaborative filtering that supports content-based filters and users. The proposed filterbots help with the problem of sparsity, however since the GroupLens predictions still use a collaborative filtering approach, new users, and hence, new filterbots, still suffer from the early rater problem. Claypool et al. (1999) proposed a similar approach that combined collaborative filtering with content-based filtering techniques and had shown to successfully mitigate most shortcomings, but scalability. Breese et al. (1998) identified two major classes of prediction algorithms - *memory-based* and *model-based*. Memory-based methods are simpler, but computationally expensive and cannot provide explanations of predictions or further insight into the data. For model-based algorithms, the model offers an intuitive rationale for recommendations making assumptions more explicit.

A method, called *personality diagnosis* was proposed by Pennock and Horvitz (2000). For large amounts of data, a straightforward application of personality diagnosis however, suffers from the same time and space complexity concerns as memory –based methods. Larsen (1999) presents a new personalization-centric concept – the idea of linking all customer touch points to a single database that proved to be a highly successful business practice.

An alternative classification of recommendation methods was proposed by Karypis (2001), where he presents a class of item-based recommendation techniques as opposed to the user-based collaborative filtering to address the scalability issues. Some recent studies have considered the use of association rule mining in recommender systems. However, there has been a little focus on the impact of factors such as the support threshold or the size of user history on the effectiveness of recommendations (Mobasher et al. 2001). Another data mining approach, the "mixed-effects" model for recommender systems was applied, using a Bayesian methodology with intention to address the cold-start issue, scalability and sparse ratings problem (Condliff et al. 1999).

O'Connor (2001) pointed out the area where rules-based personalization outperforms other techniques – with complex transactions, in which the seller has clear, structured, and well-defined practices in identifying which customers it wants to do business with and how it needs to conduct these relationships. Aggarwal (1996) studied the problem of on-line mining of customer profiles specified with association rules. Adomavicius and Tuzhilin (2001) also presented a framework for building behavioral profiles of individual users. One point becomes persistently clear from all the above discussion – the effective personalization has to be domain specific. Therefore, a profound domain analysis is very important. Our research paper focuses on personalization solutions for online customer care, thus we have investigated the specific aspects of this domain.

Personalization Approach

In this section we propose a personalization approach for an online customer care application and present the selected combination of personalization methods.

An online customer care application of a large telecommunications company that can be used in various points of contact with the customer (such as Call Centers, Internet web site, specialized retail stores) should be able to successfully apply personalization technique to two types of customers: existing and new. An existing customer is the one who has already purchased a plan, and hence has his data and at least a partial profile in the system. A new customer is the one who still has to choose a subscription plan and a phone, and for whom no profile is stored in the system. In our opinion, these two customer types require a different personalization treatment that would allow maximizing the quality of recommendations.

Our personalization approach is to create explicit customer profiles using hybrid data mining approaches based on domain models and customer data implicitly generated from current or previous customer's information and behavior. Our approach is to incorporate rule-based analysis following the solid customer profile model developed by Adomavicius and Tuzhilin (2001). A complete customer profile model consists of two parts: factual and behavioral. The factual profile contains information, such as income, age, job that the personalization system obtained from the customer's factual data. The factual profile also can contain information derived from the transactional data, such as "the customer tends to purchase only universal phone accessories" or "the customer biggest purchase was $300". A behavioral profile is usually derived from transactional data and models the customer's actions, for example, "when Josh buys a universal accessory, he also buys a complimentary phone-specific accessory". Collaborative filtering is the dominant technology for developing insight and providing recommendations. However, having analyzed the personalization issues relevant for our domain,

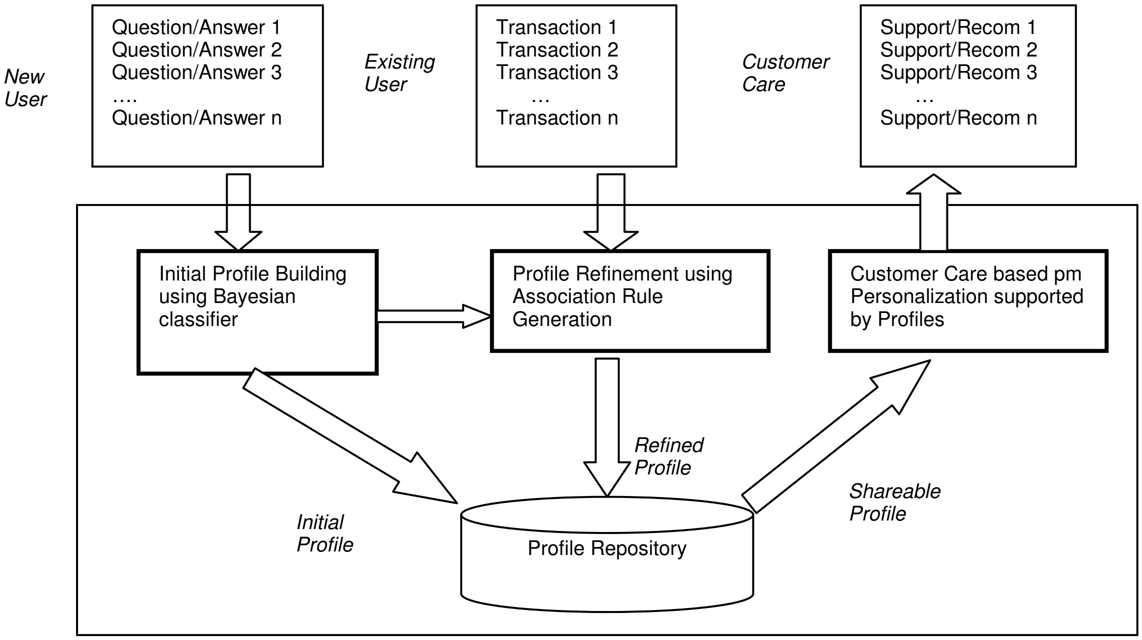

Figure 1. Architecture of Hybrid Personalization System

we have come to the conclusion that it would be ineffective for our domain. Therefore we will use Bayesian classification methods complemented by rules-based analysis that would allow action based on both customer and non-customer objectives, such as reducing over-stocked items or recommending geographic-specific subscription plan. Our recommendation model includes three components (Figure 1):

Initial profile: For a new customer that does not yet have any profile created for him, a personalization engine should be able to recommend a phone model and a subscription plan. For this purpose we will use *naïve Bayesian classifier* that can, based on, for example, customer's lifestyle, and wireless needs, match his class label, from where recommendations can be driven.

Refined profile: For an existing customer, to maintain and update the profile, we can use the data generated by *Bayesian classifier* and apply *association rules*, discovered for the class where customer belongs to. Thus we can create an accurate customer profile and, based on the profile, recommend phone-specific accessories and universal accessories. For customer retention, a successful personalization engine can also suggest an adjusted subscription plan that would better meet customer's needs.

Sharable profile: Once the customer profile has been refined with the other customer transaction-based association rules, and updated with the customer's own transaction rules, both factual and behavioral part of the customer profile are complete and it becomes a sharable profile. Now we can use this customer's profile for both generating effective recommendations this and other existing customers and refining profiles of new customers.

The following step-by-step profile enhancement and updating generates the personalization model:

1. Use *Bayesian classifier* to identify or predict a class label for a new customer (use to identify both Phone Type class label Gk and Plan Type class label Ci). On the first step we have determined which class a new customer belongs to. In our system, a data repository Di, i=1,.., m is associated with each class Ci, and contains a list of transactions, completed by the members of the class Ci. Now we describe the proposed personalization model in more detail. The *naïve Bayesian classifier* makes assumption of class conditional independence, that is, given the class label of a sample, the values of the attributes are conditionally independent of one another. For our domain the assumption holds true, and therefore the *naïve Bayesian classifier* is the most accurate in comparison with all other classifiers (Han and Kamber 2001). Once the new customer has purchased a plan and a phone, it is time to create his/her profile, which subsequently can be used to provide better recommendations to that customer, as well as to derive improved marketing solutions.

2. In the second step, we use association rules to model individual customer's behavior. Rule discovery methods are applied individually to every customer's data. To discover association rules that describe the behavior of individual customers, we use *Apriori* algorithm. Once the class labels are determined by the *Bayesian classifier* in the first step, we apply association rules generated for the existing customers of the selected Plan Type and Phone Type classes to recommend accessories/special promotions to the new customer. Since the rules are mined only for 2 respective transaction data repositories, this kind of

clustering enforces a better scalability and thus improved real-time performance of the system. Because data mining methods discover rules for each customer individually, the selected methods work well for applications containing many transactions for each customer, such as credit card, online browsing, etc. In our application domain, it is also possible to obtain a payment transaction pattern and thus address fraud issues.

3. The discovered strong association rules can be used to further recommend accessories and plan adjustments to that customer. This step completes both factual and behavioral parts of customer profile and is now becomes a *sharable* profile that can be used by the system to mine rules for and update other customers' profiles. This step completes the behavioral part of the customer profile which can now be shared by other customers.

Case Study: SprintPCS Customer Care

To demonstrate our implementation model and approach, we will use the data that belongs to a large telecommunications company, namely SprintPCS.

According to the company's annual report, the estimated number of customers as of the year 2002 is 26 million nation-wide. SprintPCS products include various phone models (around 300), phone-specific accessories (around 6000), a number of universal accessories (around 200)[1], and finally a selection of customized phone plans, targeted for various age and social groups. Our data model is an approximation of a real-life data model, and is based on information collected about SprintPCS customer care domain. It includes five classes Customer, Phones, Accessories, PlanType and PlanItemized.

Customer
a. lifestyle: {family, business_associates, friends, combo}
b. wireless needs: {socially_connected, secure, max_efficiency}
c. region: {determined by a zip code}
d. technology: {standard, latest}

Phones{Samsung, Nokia, …}
Accessories {standard_battery, vehicle_power_adapter, leather_case, hands_free_car_kit, desktop_charger, travel_charger, wireless_web_connection_kit}
PlanType {standard, total_digital_connections, family}
PlanItemized {callerId, long_distance, call_waiting, operator_services, 3way_calling, call_forwarding, voicemail, directory_assistance}

The Customer class has four attributes, each of which has several possible parameters that are meant to identify a customer's potential preferences for a phone model and plan type. The Phones class contains a list of available phone models, while the Accessories class contains the available phone accessory types. The PlanType class shows three major categories of phone plans, and PlanItemized has a list of services and features that might or might not have a particular phone plan. In this data model customer's lifestyle, wireless needs and region will influence a service plan type class selection, a phone class will be selected based on wireless needs and technology.

Let us assume there is a new customer who intends to sign up for a PCS service that is to purchase a phone and a respective phone plan.

In the first step, we have to determine customer's location to validate that the company provides coverage for that area. Next step would be to define customers' needs, or in other words, what type of customer he/she is – a business man, or family member, or a teenager who has a lot of friends to chat with, or a combination of these.

Customer lifestyle: "family"
Customer wireless needs: "max_efficiency"
Customer region: "44056"
Customer technology:"standard"
____
Customer phone range:"Samsung SPH-I300, Sanyo SCP-6000". Sanvo SCP-5750"

Figure 2. Initial customer profile

Customer lifestyle: "family"
Customer wireless needs: "max_efficiency"
Customer region: "44056"
Customer technology:"standard"
____
Customer selected phone:"Samsung SPH-I300"
____
Customer plan type:"total_digital_connections-8000minutes"

Figure 3. Refined customer profile

And final step would be to find out in what kind of technology the new customer is interested, which also determines the price of the phone model to be suggested.

Customer Profiling

Once all the essential primary data is collected we can apply the initial Bayesian classification to determine the new customer's class labels for a plan type and phone model. Once the plan type is determined, we can further apply Bayesian classifier to determine phone model type to be recommended to the customer. Also to ensure satisfaction and precision of the choice, at least three most probable phone models are displayed. This way, the recommendation engine can also learn more about this customer from the choice preference that he makes.

At this stage our system would have created the *initial* customer profile (Figure 2), based on the customer factual data only.

[1] These numbers are a result of our own observations.

Customer name: “John Doe”
Customer lifestyle: “family”
Customer wireless needs: “max_efficiency”
Customer region: “44056”
Customer technology:”standard”
____
Customer selected phone:”Samsung SPH-I300”
____
Customer plan type:”total_digital_connections-8000minutes”
____
Rules:
{if purchases long _distance, is likely to purchase operator_services}
{if purchases vehicle_power_adaptor, is likely to purchase hands_free_car_kit}

Figure 4. Sharable customer profile

Note that this profile still has a *range* of possible phones and plans that are likely to be selected by the customer. At this point we collect customer’s input about his selection and update the profile correspondingly.

Once we have classified the customer’s into specific classes based on his selection (which in case of a returning customer can be a completed transaction), we can apply *Apriori* algorithm to mine for the association rules discovered for the members of those classes and update customer’s profile with those rules, thus creating a *refined* customer profile that can look like Figure 4.

Once customer has completed one or more transactions that would include accessories and services, we can use *Apriori* algorithm and validate the previously discovered rules for that customer as well as update customer’s behavioral profile with the rules discovered from his own transactions. The discovered strong association rules can be used to further recommend accessories and plan adjustments to that customer. This step completes both factual and behavioral parts of customer profile and is now becomes a *sharable* profile that can be used by the system to mine rules for and update other customers’ profiles. After these steps are completed, the recommendation engine will display suggestions to the customer. This step-by-step profile enhancement and updating will guarantee that all customer profiles are accurate and up to date.

Previous research shows that for a given set of task-relevant data, the data-mining process may uncover thousands of rules, many of which are uninteresting for the marketing expert. Therefore, the next step would be to apply constraint-based mining, where mining is performed under the guidance of various kinds of provided constraints. These constraints can include knowledge type constraint, data constraint, dimension constraint, interestingness and rule constraints.

Comparative Analysis

Based on the previously conducted research and experimental results we have drawn a chart (Table 1) that provides a complete picture of various personalization approaches vs. limitations and shortcomings.

Scalability: While most of the today’s algorithms are able to process tens of thousands of customers and product items in real time, but the demands of modern e-commerce systems are to process tens of millions of items and customers.

Quality of recommendations: This is measured in terms of both coverage and accuracy (precision) of the produced recommendations.

Sparsity: If the number of items far exceeds what any individual can hope to absorb, the matrices containing the ratings of all items are very sparse.

Real-time performance: This factor is closely related to the problem of scalability.

The early rater issue: This has to do with providing a prediction for an item when it first appears, since there are no user ratings on which to base the predictions. Similarly, even an established system will provide poor predictions for each and every new user that enters the system.

Synonymy: This can worsen the quality of recommendations in those information domains where very similar items have different names, and the similarity cannot be recognized by the system, while evaluating customer ratings on the similar items.

The first conclusion that can be drawn from the chart is that a hybrid approach that combines several methods is more efficient than a straightforward approach. Secondly, each approach usually has its best performance area. All the issues listed in the chart might not be relevant for a particular problem domain. For example, for online customer care domain only scalability, quality of predictions, and real-time performance are the important requirements, which we have addressed and fulfilled.

Conclusions

In this paper we have conducted a profound comparative analysis of the existing personalization techniques and evaluated that a hybrid personalization approach that combines several methods is more efficient than a single approach. Since an effective personalization technique has to be customized to meet the specific needs of every particular domain and deliver quality recommendations and thus serve its purpose. Our approach based on the combination of Bayesian classification methods with association rule mining is used to model individual customer’s behavior. While Bayesian classifier can be successfully used to create effective customer profiles, rules-based analysis would allow action targeted for both customer and non-customer objectives, such as reducing over-stocked items. In this paper, we demonstrate through a case study that our recommendation system is capable of producing accurate recommendations for the online customer care problem domain.

	Scalability	**Quality of predictions**	**Sparsity**	**Early rater issue**	**Synonymy**	**Privacy**	**Real-time performance**
Pure Collaborative Filtering (Mobasher et al. 2001)	-	+/-	-	-	-	N/A	-
Pure Clickstream Analysis (Andersen et al. 2000)	-	+/-	N/A	+	N/A	-	+/-
Clustering and collaborative filtering (O'Connor and Herlocker 1999)	+	-	+/-	-	N/A	N/A	+/-
Clickstream analysis using subsessions (Breese et al. 1998)	-/+	N/A	N/A	+	N/A	N/A	+
Content-based filtering agents in collaborative filtering (hybrid approach) (Claypool et al. 1999)	+/-	+	+	-	N/A	N/A	+/-
Content-based and collaborative filters (non-hybrid approach) 6. (Sarwar et al. 1998)	-	+	+	+	-	N/A	+/-
Item-based Top-*N* Recommendation algorithms (Karypis 2001)	+	+	N/A	-	N/A	N/A	+
Bayesian Mixed-Effects Model (Condliff et al. 1999)	+	+/-	+	+	N/A	N/A	+
Rules-based personalization (O'Connor 2001)	-	+	+/-	-	-	N/A	-
Pure Data-mining methods (Mobasher et al. / Adomavicius and Tuzhilin 2001),	+/-	+/-	+/-	-	-	-	+/-
Our personalization model	+	+	N/A	N/A	N/A	N/A	+

("N/A": Not Addressed, "+" : issue is resolved, "-" : issue is not resolved , "+/-": issue is partially resolved)

Table 1. Comparative analysis of personalization approaches

References

Agrawal, R. Mannila, H. Srikant, R. Toivonen, H. Verkamo. 1996. Fast discovery of association rules. In Advances in Knowledge Discovery and Data Mining, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.). AAAI Press, Melo Park, CA.

Andersen, J., Larsen, R. S., Giversen, A., Pedersen, T.B., Jensen, A.H., Skyt, J. 2000. Analyzing Clickstreams Using Subsessions. In Proceedings of the 3rd International Workshop on Data Warehousing and OLAP, pp. 25 – 32.

Adomavicius, G., Tuzhilin, A. 2001. Using Data Mining Methods to Build Customer Profiles. IEEE Computer 34(2): 74-82

Adomavicius, G., Tuzhilin, A. 2001. Expert-Driven Validation of Rule-Based User Models in Personalization Applications. Data Mining and Knowledge Discovery 5(1/2): 33-58

O'Connor, A. 2001. Personalization Coming Full Circle, Part 2.

Badrul M. Sarwar, Joseph A. Konstan, 1998. Al Borchers, Jon Herlocker, Brad Miller, John Riedl, Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System. In Proceedings of the ACM Conference on Computer Supported Cooperative Work.

Breese, J.S., Heckerman, D., Kadie C. 1998. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In Fourteenth Conference on Uncertainty in Artificial Intelligence, November.

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M. 1999. Combining Content-Based and Collaborative Filters in an Online Newspaper. ACM SIGIR Workshop on Recommender Systems, Berkeley.

Condliff, M.K., Lewis, D.D., Madigan, D., Posse C. 1999. Bayesian Mixed-Effects Models for Recommender Systems. In ACM SIGIR Workshop on Recommender Systems: Algorithms and Evaluation.

Larsen, S. 1999. Developing the Personal-Centric Enterprise through Collaborative Filtering and Rules-Based Technologies. A CRM Project.

Mobasher, B., Dai, H., Luo, T., Nakagawa, M. 2001. Improving the Effectiveness of Collaborative Filtering on Anonymous Web Usage Data, In Proceedings of the IJCAI 2001 Workshop on Intelligent Techniques for Web Personalization (ITWP01), Seattle.

Mobasher, B., Dai, H., Luo, T., Nakagawa, M. 2001. Effective Personalization Based on Association Rule Discovery from Web Usage Data, In Proc. of ACM Workshop on Web Information and Data Management (WIDM) pp. 103-112.

O'Connor, M., Herlocker, J., 1999. Clustering Items for Collaborative Filtering, ACM SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation

George Karypis, 2001. Evaluation of Item-Based Top-*N* Recommendation Algorithms. In Proceedings of CIKM.

Pennock, D.M., Horvitz, E. 2000. Collaborative Filtering by Personality Diagnosis: A Hybrid Memory – and Model-Based Approach. In Uncertainty in Artificial Intelligence, pages 473 – 480.

Han, J., Kamber, M. 2001. Data Mining: Concepts and Techniques, pp. 298-300