Personalization for Web-based Book Shop System using Hybrid Data Mining Approaches

Thazin Hlaing, Pearl
Computer University, Pathein
gaung88@gmail.com ; pearl417@gmail.com

Abstract

Personalization is a new system development approach for designing information systems that change configurations based on each user's needs and preferences. If it is possible to recommend products to customers' liking at the time they are visiting the specific web site, it would reduce the hassle customers experience in searching for products from a large information base.

An effective personalization technique has to be customized to meet the specific needs of every particular domain and deliver quality recommendations. This paper presents a hybrid personalization system for web-based book-shop system. It combines the Bayesian classification method with association rule mining to model individual customer's behavior. While Bayesian classifier is for effective customer profiles, Association rule presents item to item association based on user transactions.

Keywords: Recommender system, hybrid data mining approach, Bayesian classifier, Association rule mining

1. Introduction

The growth of internet has begun to change many people reading to book. The on-line book portals enable the user to download the book into their computer from the book web sites according to their likes. In observing the user, there are various users who want to read to specific book classes. There are a variety of new books and so many books are already stored. There is a difficulty of searching which kinds of book he/she likes.

Recommending the users based on his/her attributes helps the user's choosing patterns and automatically personalizes the interface presentation so as to aid the use in making reading or downloading. Today, users have more choice to read to books than ever. They are more aware of the possibilities and demanding of personal attention. It is now becoming increasingly important for a company to build a strong relationship with its user. This is where personalization technology steps in. Users can be classified into "classes" based on past behavior as well as predictions of future behavior.

Though, there exists 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. This hybrid data mining approach proves that an effective personalization technique has to be customized to meet the specific needs of every particular domain. Then only it would be able to deliver quality recommendations and thus serve its purpose. While Bayesian classifier can be successfully used to create effective customer profiles, rules-based analysis would allow to find associations between recommend items.

This paper is organized as follows. Section 1 is the introduction, section 2 is related work. About personalization process is presented in section 3. In section 4, proposed system design, Bayesian analysis, association rules, process flow and database design are presented. Section 5 is the system implementation for recommending items using hybrid data mining approach. Section 6 is the conclusion of the system.

2. Related Work

Automatic personalization implies that the user profiles are created potentially updated, automatically by the system with minimal explicit control by the user.

Traditional approach to automatic personalization has included content-based, collaborative, and rule-based filtering systems. Each of these approaches is distinguished by the specific types data collected to construct user profiles, and by the specific types of algorithmic approach used to provide personalize content. 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)[1]. 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) [7].

GroupLens (Sarwar et al. 1998) [2] 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 have used a collaborative filtering approach, new users, and hence, new filterbots, still suffer from the early rater problem. In Claypool et al. (1999) [5] a similar approach is proposed that combined collaborative filtering with content-based filtering techniques and had shown to successfully mitigate most shortcomings, but scalability. In Breese et al. (1998) [4] two major classes of prediction algorithms are identified; 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.

3. Personalization

Personalization is a process of gathering and storing information about site visitors, analyzing the information, and based on the analysis, delivering the right information to each visitor at the right time. [10] Personalization is the combined use of technology and customer information to tailor electronic commerce interactions between a business and each individual customer. Using the information either previously obtained or provided in real-time about the customer and other customers, the exchange between the parties is altered to fit that customer's stated needs so that the transaction requires less time and delivers a product best suited to that customer. Personalization is the capability to customize communication based on knowledge preferences and behaviors at the time of interaction. Personalization is about building customer loyalty by building a meaningful one-to-one relationship; by understanding the needs of each individual and helping satisfy a goal that efficiently and knowledgeably addresses each individual's need in a given context. Personalization system falls into three basic categories:

- Rule based Personalization System,
- Content based Filtering System, and
- Collaborative based Filtering System.

Personalization is one of the keys for the success of web services. Web Personalization aims to adapt the Website according to the user's activity or interests. The information architecture components for personalization comes the three areas of personalization context, books and users.

- Personalization Context: The system has certain rules in which association of book type and its user profile that determine how personalization happens.
- Book: Likewise, book is profiled, based on a set of attributes and assigned specific values of each book (such as name of the book, author name, book category, publisher, book price and so on.
- Users: User profiles represent their interest behaviors. Specific values for a profile are determined by the set of defined attributes and possible values for each attribute (such as user name, address, occupation, age, gender and so on.)

4. Proposed System

This paper presents a personalization system for an online user application and presents the selected combination of personalization methods. In this system, there are two types of users: existing and new. An existing user is the one who has already chosen a book, and hence has his data and at least a partial profile (transaction data) in the system. A new user is the one who still has to choose a book, and for whom no profile (transaction data) is stored in the system. These two user types require a different personalization treatment that would allow maximizing the quality of recommendations. This personalization approach is to create explicit user profiles using data mining approaches based on domain models and user data implicitly generated from current or previous user's information and behavior (sales data).

It is to incorporate rule-based analysis following the solid customer profile model. A complete customer profile model consists of two parts: factual and behavioral. The factual profile contains information, such as income, age, occupation that the personalization system obtained from the customer's factual data. The

behavior profile includes information derived from user transactional data.

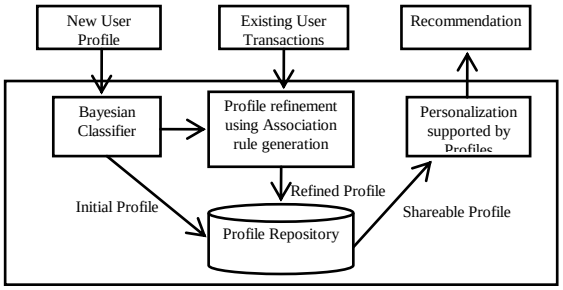

Figure 1: System Overview

Figure 1 presents the overview process of proposed system. It shows the process of both new user and old user.

4.1 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 book types. Bayesian classifier to determine book types to be recommended to the customer.

- **Initial Profile**: For a new customer that does not yet have any profile created for him, a personalization engine should be able to recommend the book.
- **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.
- **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. It can be used for both generating effective recommendations.

4.2 Bayesian Analysis

Bayesian classifiers are statistical classifier. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Bayesian classification is based on Bayesian theorem. [3]

Bayesian Theorem

Given training data X, posteriori probability of a hypothesis H, P(H|X), follows the Bayes' theorem

P (H/X) =P(X/H) P (H) / P(X),

- P(H) (prior probability), the initial probability
- P(X): probability that sample data is observed
- P(X|H) (posteriori probability), the probability of observing the sample X, given that the hypothesis holds

Informally, this can be written as

posteriori = likelihood x prior/evidence

Predicts X belongs to C_i if the probability $P(C_i|X)$ is the highest among all the $P(C_k|X)$ for all the k classes. [6] Advantage of naïve Bayes classifier is that it requires a small amount of training data to estimate the classification. Given training data X, posteriori probability of a hypothesis H, P(H|X) follows the Bayes' theorem –

$$P(H|X)=\frac{P(X|H)P(H)}{P(X)}$$

Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X|C_i)$. In order to reduce computation in evaluation $P(X|C_i)$, the naïve assumption of class conditional independence is made .This presumes that the values of the attributes are conditionally independent of one another, given the class Label of the tuple.Thus

$$P(X|C_i)=\prod_{k=1}^{n} P(x_k|C_i)$$

$$= P(x_1|C_i)\times P(x_2|C_i)\times \ldots\ldots \times P(x_n|C_i)$$

4.3. Association Rules

Association rule mining finds interesting association among a large set of data items. [3] Association rules are considered interesting if they satisfy both a minimum support and a minimum confidence threshold. A more formal definition is as follows: Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items. Let D be a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is implication of the form $A \rightarrow B$, where $A \subset I$, B C I and $A \cap B = \varnothing$. The rule $A \rightarrow B$ holds in the transaction set D with support ***s***, where ***s*** is the percentage of transactions in D that contain $A \cup B$. This is taken to be the probability, P ($A \cup B$). The rule $A \rightarrow$ B has confidence ***c*** in the transaction set D if ***c*** is the percentage of transactions in D containing A that also contain B. [8, 9] This is taken to be the conditional probability, P(B|A). That is,

Support ($A \rightarrow B$) = P ($A \cup B$)

Confidence ($A \rightarrow B$) = P (A | B)

4.4 Process Flow of the System

Figure 2 and 3 present the process flow of new customer and old customer.

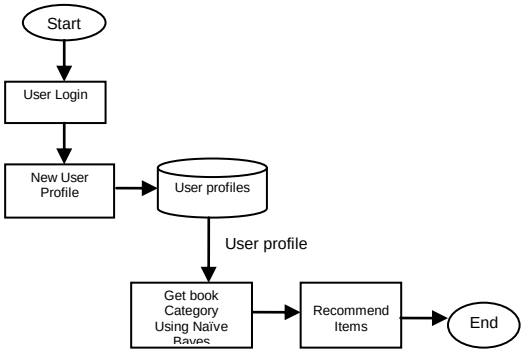

Figure 2: Process Flow of personalization(new user)

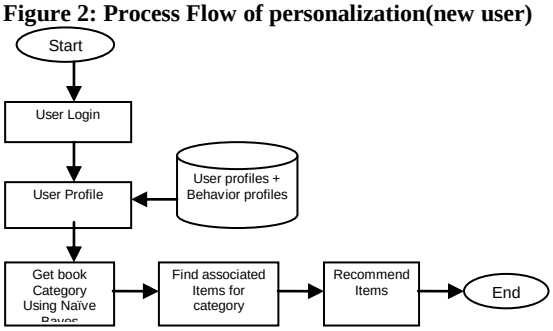

Figure 3: Process Flow of personalization (old user)

In Figure 2, when the new user (user who does not buy a book) logins, customer factual profile is used to get the class of the user. User is recommended with the books classified by Bayesian classifier. In Figure 3, when the old user (user has transactions) logins, customer factual profile is built as in new user. Then class of the user is specified and user is recommended based on the results of association rules between user transaction and categories of Bayesian classifier.

4.5 Database Design

This system stores, customer profiles and transactions. It also contains book information. Transaction data is stored in two table Sales table and sales detail table. Sales table contains header information like sales date, customer name and sales detail table contains the detailed list of the books that the customer bought. In this system book id is standardized to two characters of its category and numerical values. Database design is shown in Figure 4.

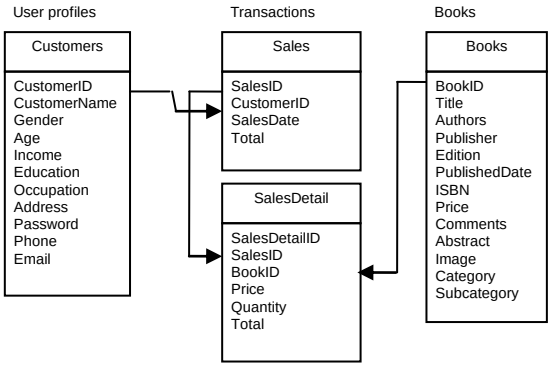

Figure 4: Database Design

5. System Implementation

This system is an implementation for personalization of book store recommender system. Customers are recommended based on factual profile (their attributes) and behavior profile (sales transactions).

Bayesian classification algorithm is combined with association rule to get the recommendation sets. Bayesian classification algorithm is used to estimate class label for customers, and the association rule mining algorithm is applied to generate associated book items for existing customer by using Transaction table. It is implemented as web based program providing books related with IT field. Following tables show customers table, books table and transactions table.

Table 1: Customer Table

Cust ID	Cust Name	Gender	Age	Income	Educat ion	Occupation
C1	Aye Aye Myint	F	22	Medium	M.C.Sc	Programmer
C2	Aung Kyaw	M	25	High	B.C.Sc	Programmer
C3	Cherry	F	21	High	B.C.Sc	Student
C4	Ei Ei	F	30	High	B.E	Engineer
C5	Su Mon	F	25	Low	Eco	Accountant
C6	Ko Ko	M	21	Low	B.C.Sc	Student
C7	Linn Linn	M	27	High	M.E	Programmer

Table 2: Book Table

Book ID	Title	Category	Sub Category

AI01	Fuzzy Logic and Expert Systems Applications	Artificial Intelligence	Fuzzy Logic and sets
AI02	Fuzzy Expert System and Fuzzy Reasoning	Artificial Intelligence	Expert System
BS01	Request for Proposal: A Guide to Effective RFP Development	Business	General
DB01	MS SQL Server 2005 Administrator's Companion	DBMS	MS SQL Server
EX01	The A+ Certification and PC Repair Handbook	EXAM	A+
EX02	CCNA Official Exam Certification Library	EXAM	CCNA
PG01	Pro Visual C++ and.NET 2.0 Platform	Programming	ASP .NET C#

Table 3: Transaction Table

TID	CustID	Items
T1	C1	AI01,PG01,AI02
T2	C4	EX01,EX02
T3	C1	DB01
T4	C2	PG01
T5	C7	PG01, EX01,EX02
T6	C3	AI02,PG01,AI01
T7	C5	BS01
T8	C6	PG01,DB01,AI02

Following is the sample recommendation for old customer C2.

X=(Gender=M,Age=25,Occupation=Programmer, Income=High,Education=B.C.Sc)

By Bayesian computation,

P(C2|buy='AI')P(Category='AI')=0.005568

P(C2|buy='Business')P(Category='Business')=0

P(C2|buy='DBMS')P(Category='DBMS')=0

P(C2|buy='Exam')P(Category='Exam')=0

P(C2|buy='Programming')P(Category= 'Programming')=0.0300672

Artificial Intelligence and Programming are maximum probabilities. This customer C2 has already bought programming PG01. According to association rules, suppose minimum support = 20% and minimum conf = 50%.

PG01 → AI01 = 2,min_ sup = 2 / 8 = 25%, conf = 2 / 5 = 40%

PG01 → AI02 = 3, sup = 3 / 8 = 38%, conf = 3 / 5= 60%

According to association rule and Bayesian probability, he will be recommended AI02 (Fuzzy Expert System and Fuzzy Reasoning). Moreover, system also recommend available books list for this customer.

For old user, Bayesian probability is computed for class label. Then Association rule mining is applied to find the associated items with user transaction based on class label of Bayesian. Recommending old user is shown in Figure 5.

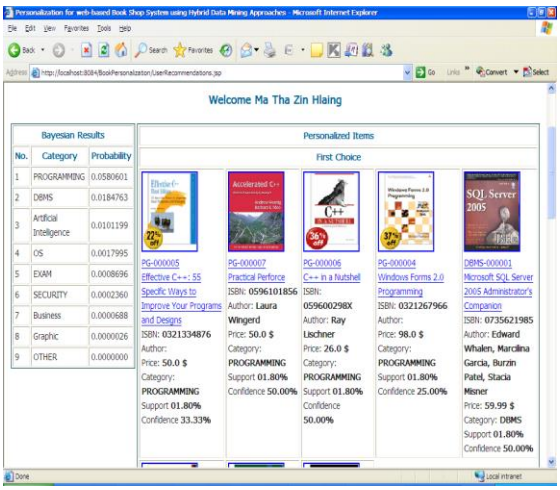

Figure 5: Recommended Page for old User

For the new user, only Bayesian probability is computed since there is no transaction for the new user. Recommender Page for new user is shown in Figure 6.

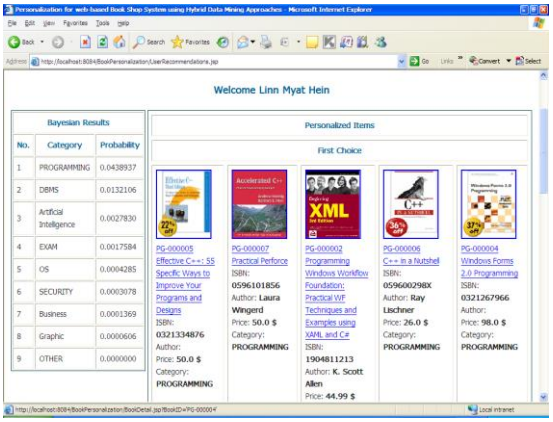

Figure 6: Recommended Page for New User

5.1 Classifier Accuracy

Estimating classifier accuracy is important since it determines to evaluate how accurately a given classifier will be labeled future data, data on which the classifier has not been trained. Accuracy estimates also help in the comparison of different classifiers. A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with

training set used to build the model and test set used to validate it.

Accuracy ratio: the percentage of test set samples that are correctly classified by the model

The Sensitivity and specificity measures can be used to determine the accuracy measures. Precision may also be used to access the percentage of samples labeled as for example, "yes" that actually are "yes" samples. These measures are defined as:

Sensitivity = t_pos / pos
Specificity = t_neg / neg
Precision = t_pos / (t_pos + f_pos)

Where,

t_pos = the number of true positives ("yes" samples that were correctly classified as such),

pos = the number of positive ("yes") samples

t_neg = the number of true negative ("no" samples that were correctly classified as such)

neg = the number of negative samples

f_pos = number of false positive ("no" samples that were incorrectly labeled as "yes")

$$\text{accuracy} = \text{sensitivity}\,\frac{\text{pos}}{(\text{pos}+\text{neg})} + \text{specificity}\,\frac{\text{neg}}{(\text{pos}+\text{neg})}$$

In this paper, accuracy measure is used to measure the performance of the system. There are 1500 transactions and 147 user profiles to test the system. According to the experimental results, combination of Bayesian classifier and association rule mining algorithm got 85.7% of accuracy for the recommender system while Bayesian classifier got 76.12% of accuracy.

6. Conclusion

In this system, an effective personalization technique has to be customized to meet the specific needs of every particular domain and deliver quality recommendations. In this system, while Bayesian classifier can be successfully used to create effective customer profiles, rules based analysis find item association based on user transaction. So our recommendation system is capable of producing accurate recommendations for the online customer care application and a hybrid personalization approach is more efficient than a single approach.

7. References

[1] B. Mobasher, H. Dai and T. Luo, Nakagawa, M. "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, 2001.

[2] B.M. Sarwar and A. K. Joseph, "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, 1998.

[3] J. Han, "Data Mining: Concepts and Techniques", Second Edition, ISBN 1-55860-489-8.

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

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

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

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

[8] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94), pages 487–499, Santiago, Chile, Sept. 1994.

[9] R. Agrawal, T. Imielinski, T. and A. Swami, "Mining association rules between sets of items in large databases", Proceedings of the ACM SIGMOD, 1993, pp. 207-216.

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