Recommendation System Using Web Data Mining

Mihir A. Mishra
Department of Computer Engineering
GIDC Degree Engineering College, Navsari, India

Abstract- *A recommendation system is an enabling mechanism to overcome information overload occurred when shopping in an Internet marketplace. It uses log files to get the information about users accessing patterns. Based on this information the products are recommended to the customers. This creates a user friendly environment. Collaborative filtering has been known to be one of the most successful recommendation methods, but its application to e-commerce has exposed well-known limitation such as scarcity and scalability, which would led to poor recommendation [1]. The method in this seminar suggests a personalized recommendation methodology by which we are able to get further effectiveness and quality of recommendations when applied to an Internet shopping mall. The suggested methodology is based on a variety of data mining techniques such as web usage mining, decision tree induction, association rule mining and product taxonomy.*

Keywords- E-commerce, Data Mining, Web Mining, Recommendation System, Clustering, Decision Tree

I. INTRODUCTION

E-commerce is growing at rapid pace and keeping pace with rapid growth is a challenge to both companies and customers. Therefore the need for new marketing strategies such as one to one marketing and customer relationship management has been stressed both from researches and from practical affairs [1]. One solution to these strategies is personalized recommendation system that helps customers find the products they would like to purchase by producing a list of recommended products for each given customer.

Collaborative filtering is the best known recommendation system until now. It identifies the customer whose interests are similar to those of a given customer and recommends products neighbors of a given customer has liked [1]. However these techniques show two limitations. The first is related to sparsity. The number of rating already obtained is very small compared to the number of ratings that needed to be predicted because collaborative filtering requires explicit non-binary user rating for similar products. As a result collaborative filtering cannot accurately compute neighborhood and identify products to be recommended. The second is related to scalability. Algorithm that needs to find neighborhood requires complex calculation. With millions of customers and products of real world situation existing collaborative filtering based recommendation system suffer serious scalability problem.

Recent studies have suggested web usage mining as an enabler to overcome the problems associated with collaborative filtering since it will reduce the need for obtaining subjective user ratings or registration based personal preferences [1]. Click stream in Internet shopping mall provides information essential to understand shopping patterns or repurchase behaviors of the customers such as what product they see, what product they add to the shopping cart and what they buy. Mining association rules from Click stream provides rich and interesting relationship or associations among products, which are used in characterizing the appeal of individual products, compared to the mining association from purchase records [4].

Introduction to Data Mining

Data mining is the process of extracting useful information from huge amount of data. Alternative names are Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, business intelligence, etc [6].

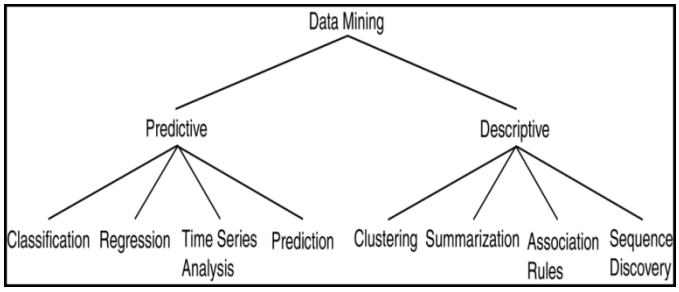

Fig 1 Data Mining Model and Task

Data Mining Techniques

1) Association Rule Mining

Association rules are typically mined by following a two-phase strategy as follows: Phase-1: Frequent item set mining (FIM) algorithm are applied on the dataset to retrieve the entire frequent item set along with their supports.Phase-2: In this phase, the information output by phase 1 is used to form interesting association among the frequent patterns.

Phase 2 does not need access the dataset and therefore, it is relatively straightforward.

2) Supervised Learning

Classification techniques are used to classify data records into one among a set of predefined classes. They work by constructing a model of a training dataset consisting of example records with known class label Classification is a two-phase approach as follows: Phase-1: Model Construction: During this phase, the labeled examples are analyzed and a model is built and stored on disk. Phase-2: Classification: During the phase, the model stored on the disk is loaded into the main memory and used to classify new unlabeled instance [6].

3) Cluster Analysis

Clustering is the task of organizing data into groups (known as clusters) such that the data object that are similar to each other are out in the same cluster. It's evaluated based on the similarity of objects within each cluster and the dissimilarity of objects across clusters [6].

4) Web data mining

The web data mining is the process of applying various data mining techniques to web contents.. Three types, based on sources of data: 1)Web structure mining, 2)Web content mining, 3)Web usage mining. Web content mining is the process of applying various data mining techniques like classification, clustering on the web documents. Web structure mining uses hyperlink structure to maintain users accessing style. Web usage mining uses log file to keep the record of the user's session based on which various recommendation systems are maintained in e-commerce [7].

5) Data warehousing

It is a technology that allows one to gather, store, and present data in a form suitable for human exploration. This involves cleaning and data integration. Although this possible using the standard relation database technology, data warehouses make the process effective and efficient.

Introduction to web mining

Web mining is the application of data mining techniques to find interesting and potentially useful knowledge from web data. It is normally expected that either the hyperlink structure of the web or the web log data or both have been used in the mining process.

There are three different types of web mining:

Web content Mining: It deals with discovering useful information from the web documents. In contrast to Web usage mining and Web structure mining, Web content mining focuses on discovering useful information or knowledge from the web page contents rather than the links. Search engines, subject directories, intelligent agents, cluster analysis and portals are employed to find what user is looking for. The content of web pages includes no machine readable semantic information [7].

Web Structure Mining: It deals with discovering and modeling the link structure of the web. Work has been carried out to model the web based on topology of the hyperlinks.This help in discovering similarity between sites or in discovering important sites for a particular topic or discipline or inn discovering web communities [7].

Web usage mining: It deals with understanding user behavior in interacting with the web or with the web site. One of the aims is to obtain information that may assist website reorganization or assist site adaptation to better suit the user. The mined data includes data logs of users interaction with the web. The logs include the web server logs, proxy server logs and browser logs. The logs include information about the referring pages, user identification, time a user spends at a site and the sequence of pages visited. Information is also collected via cookie files. While web structure mining shows that page A has a link to page B, web usage mining shows who or how many people took that link, which site they came from and where they went when they left page B [7].

The motivation behind developing recommendation system is the fast growing e-commerce application. The recommendation will maintain the users accessing pattern using log files and recommend only those products that are of users interest. This will create a user friendly environment.

The objective of the recommendation system is to suggest to the customers only those products that are of their interest based on their accessing pattern, thus reducing the time for selecting the product.

The recommendation system recommends the product to the customers based on their previous accessing style. The use of this is in stock market analysis, sensor network analysis, e-commerce application and so on.

II. THEORETICAL BACKGROUND

Recommendation System

An information filtering technology, commonly used on e-commerce Web sites that uses a collaborative filtering to present information on items and products that are likely to be of interest to the reader. In presenting the recommendations, the recommender system will use details of the registered user's profile and opinions and habits of their whole community of users and compare the information to reference characteristics to present the recommendations.

They are based on a number of technologies:

1. Information filtering: search engines
2. Machine learning: classification learning
3. Adaptive and personalized system: adaptive hypermedia
4. User modeling

The quality of recommendation has an important effect on the customer's future shopping behavior. Poor recommendation can cause two types of characteristic errors: false negatives, which are products that are not recommended, though the customer would like them, and false positives, which are products that are recommended, though the customer does not like them. In an e-commerce environment the most important errors to avoid are false positives, because these errors will lead to angry customers and thus they will be unlike to revisit the site. If we try to find customers who are likely to buy recommended products and recommend products to only them, that could be a solution to avoid false positives of poor recommendation.

Types of Recommendation System

The various types of recommendation systems are:

1. Content based Recommendation system
2. Collaborative Recommendation System
3. Cluster Models

Content based Recommendation system

The content based recommendation system characteristics are as follows:

1. Recommend items similar to those users preferred in the past.
2. User profiling is the key.
3. Items/content usually denoted by keywords.
4. Matching "user preferences" with "item characteristic works for textual information.
5. Vector Space Model widely used.

The limitation of Content based Recommendation System is:

1. Not all content is well represented by keywords, e.g. images
2. Items represented by same set of features are indistinguishable.
3. Overspecialization: unrated items not shown.
4. Users with thousands of purchases are a problem.
5. New user: No history available.
6. Shouldn't show items that are too different

Collaborative Recommendation System

The basic characteristics of Collaborative Recommendation System are:

1. Use other users recommendations (ratings) to judge item's utility
2. Key is to find users/user groups whose interests match with the current user
3. Vector Space model widely used (directions of vectors are user specified ratings)
4. More users, more ratings: better results
5. Can account for items dissimilar to the ones seen in the past too.

The limitation of collaborative recommendation system is:

1. Different users might use different scales. Possible solution: weighted ratings, i.e. deviations from average rating.
2. Finding similar users/user groups isn't very easy.
3. New user: No preferences available.
4. New item: No ratings available.
5. Demographic filtering is required.
6. Multi-criteria ratings is required.

Cluster Models

The basic characteristics of cluster models are:

1. Create clusters or groups
2. Put a customer into a category.
3. Classification simplifies the task of user matching.
4. More scalability and performance.
5. Lesser accuracy than normal collaborative filtering method.

Item to item Collaboration

The basic characteristics of item to item collaboration system are:

1. Compute similarity between item pairs.
2. Combine the similar items into recommendation list
3. Vector corresponds to an item, and directions correspond to customers who have purchased them.
4. "Similar items" table built offline.

Challenges in Recommendation System

1. Generic user models (multiple products and tasks)
2. Generic recommender systems (multiple products and tasks)
3. Distributed recommender system (users and products data are Distributed)
4. Portable recommender systems (user data stored at user side)
5. (user) Configurable recommender systems
6. Multi strategy – adapted to the user
7. Privacy protecting RS
8. Context dependent RS
9. Emotional and values aware RS
10. Trust and recommendations
11. Persuasion technologies
12. Easily deployable RS
13. Group recommendations.

Application of Recommendation System

The Recommendation systems are mainly used in:

1. Stock market analysis for prediction of shares.
2. E-commerce application in which users accessing patterns is traced based on which products are recommended to the customers

III. DESIGN AND ANALYSIS

An approach for Recommendation system

There are basically five steps involved in a Recommendation system. They are:

1. Problem Definition
2. Target Customer Selection
3. Customer Preference analysis
4. Product association analysis
5. Recommendation generation

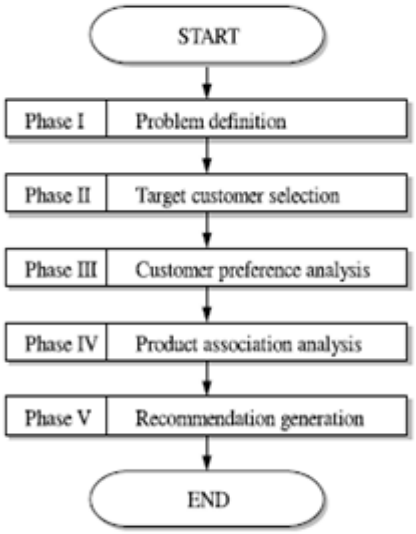

Fig 2 The overall flow of recommendation [1]

Problem Definition

In most cases, the recommendation problem in e-commerce can be classified 1] whether the system is designed for the customer as a whole or selective customers. 2] whether the objective of recommendation s to predict how much a particular customer will like a particular product, or to identify a list of product that will be of interest to a given customer. 3] Whether the recommendation is accomplished at a specific time or persistently. Most marketing campaign management systems make top-N recommendations for particular customer at the specified time. This approach only considers the recommendation problem helping selective customers find which products they would like to purchase for each of them at the specific time. Thus this approach would be useful in developing campaign management system than the collaborative management system.

Target customer Selection

Making recommendation only for customers who are likely to buy recommended product could be a solution to avoid false positives of the poor recommendation. This phase performs the tasks of selecting such customers based on decision tree induction. The decision tree induction uses both the model set and the score set generated from customer records. To generate the model and score set for the recommendation problem Rec(l,n;p,t),also needs two sets: one is the model candidate set which is a set of customers who constitute the model set and the other is the score candidate set which is a set of customers who form the score set. Let msst, pd, pl and pr be the start time of the model set, the time period for the distant past, the time period of latency , the time period of the recent past, respectively. The model candidate set is

defined as a set of customers who have purchased p or more level-l product classes between msst time and msst+pd time [1]. To make model set from model set from model candidate set, the value of independent variables in the model set are simply obtained from the attributes related to customers who belong to model candidate set. In this study a dependent variable is whether or not a customer is likely to buy new products which he/her has not yet purchased. Since the dependent variable does not exist in the customer records, it has to be generated using purchase records. The values of independent variables of the score set are also generated from records about customers who belong to the model candidate set. The decision tree assigns a class to the dependant variable of a new case, in the score set based on the values of independent variables.

Customer Preference Analysis

The methodology applies the results of analyzing preference inclination of each customer to make recommendation. For this purpose, a customer preference model is constructed based on the following three general shopping steps:

1. Click through: the click on the hyperlink and the view of the web page of the product.
2. Basket placement: the placement of the product in the shopping basket.
3. Purchase: the purchase of product-completion of a transaction.

The idea of measuring the customer's preference is simple and straightforward. The customer's preference is measured by only counting the number of occurrence of URL's mapped to the product from clickstream of the customer. If all the customers in an online store buy products only in accordance with three sequential step, then the products can be classified into four product groups such as purchased products, products placed in the basket, products clicked through, and the other product as shown in fig 3

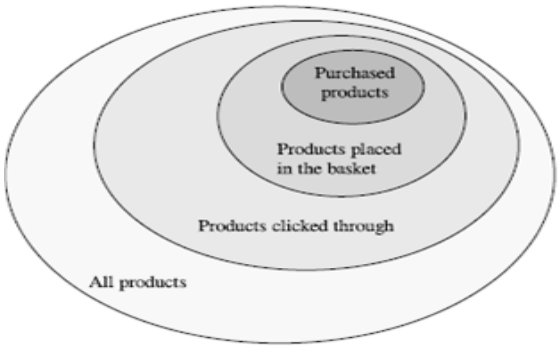

Fig 3 Classification of product according to customers shopping behavior [1]

This classification provides is-a relation between different groups such that purchased product is-a product placed in the basket, and products placed in the basket is-a click through. From this relation it is reasonable to obtain a preference order between products such that {products never clicked} < {products only clicked through} < {products only placed in the basket} < {purchased products}. The weights assigned to products placed in the basket will be higher than those placed in the basket.

Product association analysis

In this phase meaningful relationships or associations among product classes through mining association rules from the large transaction. Association rule mining is performed at level-l of the product taxonomy. To capture the e-shopper's shopping inclination more accurately, unlike the traditional usage of association rule mining, association rule from three different transaction set: purchase transaction set, basket placement transaction set and click through transaction set. The step involved in mining level-l association rule from different transaction sets are as follows:

Step 1: set the given time interval between msst time and t-1 time.
For each of the purchase transaction set, basket placement transaction set and click-through transaction set.

Step 2: Gather all the transaction made in the given time period into a single transaction in the form of <customer ID, {a set of products}>

Step 3: Find association rules among level-1 product classes according to following sub-steps:

Step 3.1: set minimum support and minimum confidence.

Step 3.2: Replace each product in transaction set with its corresponding level-l product class.

Step 3.3: Find all frequent itemsets of size 2 using Apriori or its variants

Step 3.4: Generate association rule containing a single product class in both body and head from the set of all frequent itemsets of size 2.

Next step calculates the extent to which each product class appeals to each customer from the discovered rules. This work results in building a model called product association model represented by a matrix [1].

Recommendation Generation

A Recommendation system list for a specific customer is produced by scoring each candidate product for him/her and selecting the best match. This score reflect the

degree of similarity between the customer preference and the product association. This methodology use cosine coefficient to measure the similarity the matching scores Sij between customer i and level-l product class j is computed as follows:

$$s_{ij} = \frac{\mathbf{P}_i \cdot \mathbf{A}_j}{\|\mathbf{P}_i\| \|\mathbf{A}_j\|} = \frac{\sum_{k=1}^{N} p_{ik} a_{jk}}{\sqrt{\sum_{k=1}^{N} p_{ik}^2} \sqrt{\sum_{k=1}^{N} a_{jk}^2}}$$

Equation 1

Where Pi is a row vector of the M*N customer preference matrix P, and Aj is a row vector of the N*N product association matrix A. Here, M refers the total number of customers and N refers the total number of level-l product classes. The Sij value range from 0 to 1 where more similar vectors results in bigger value. The entire product in the same product class would have identical matching scores for a given customer since the scores are computed at the level of product classes but not at the product level. From the problem definition we have to choose which of the product classes are to be recommended to the customer. The three different strategies related with such a choice:

1] Recommendation of the most frequently purchased product: This is a strategy based on the purchase history information for choosing one product per matched product class. This assumes that the more popular product implies the more buyable product. This strategy is commonly used in collaborate filtering based recommender systems for online or off-line stores.

2] Recommendation of product with the highest click-to-buy rate: This is a strategy based on customer behavior information. The click-to-buy rate measures how many click-through are converted to purchases. This strategy assumes that the more click-to-buy rate implies the increased effectiveness of marketing.

3] Recommendation of the latest product: This is a strategy based on product data. This strategy is used under assumption that many customer want fashionable item.

The decision of choosing which strategy is determined by heuristic knowledge and also depends on the application. The steps for choosing recommended products from product classes using the matching score.

Step 1: Select the choice strategy.

Step 2: Determine the number of recommended product classes, nc, such that nc<n and n/nc is an integer.
For each customer:

Step 3: select the highest scored product classes.

Step 4: Make a recommendation list which consists of n/nc product class, according to the selected choice strategy.

IV. CONCLUSION

- A recommender system main task is helping to choose products that are potentially more interesting to the user from a large set of options
- Recommender systems "personalize" the human computer interaction – make the interaction adapted to the specific needs and characteristics of the user
- Personalization is a complex topic: many factors and there is no single theory that explains all.

The characteristics of the suggested methodology are as follows

- The customer preference and other product association are automatically learned from click stream, unlike other methods which learn from purchase records only.
- In order to avoid poor recommendation that will lead to the disappointment customers, customers who are likely to buy the product are selected through decision tree induction.
- The explicit participation of the marketers and formal usage of background knowledge such as the product taxonomy are also introduced in the recommendation process.

REFERENCES

[1] Y.H cho; J.K. Kim; S.H. Kim, "A personalized recommender system based on web usage mining and decision tree induction", Expert System, vol 23, no 3, pp 329-342, 2002.

[2] J.J.M. Guervos; P.A. Castillo; B.P. Campos; J.C. Canada; F.T. Garcia; G. Ferreres, "Weblog Recommendation using association Rules" , In proc of International Conference of Web Based Communities, spain, 26-28 Feb, 2006

[3] D. Dixit; J. Gadge, "Automatic recommendation for online users using web usage mining", International Journal of Managing Information Technology, vol 3, no 3, pp 33-42, 2010

[4] H.H Inbarani; K. Thangavel, "Rough set based user profiling for web personalization", International Journal of Recent trends in Engineering, vol 2, no 1, pp 103-107, 2009

[5] J. Velasquez; H. Yasuda; T. Aoki, "Combining the web content and usage mining to understand the visitors behavior in a website" , International conference on Data mining, Melbourne, Nov 19-23, 2003.

[6] Jiawei Han, Micheline kamber, jian Pei "Data mining concepts and techniques" -The Morgan Kaufmann series in Data Management Systems

[7] Introduction to Data Mining with case studies, 2nd edition, G.K Gupta.