Data Mining

For

Recommendation System in e-commerce

By

EunJoeng Cha (683998)

To fulfill the partial requirement of Master degree in Computer Science at University of Bridgeport.

August 30, 2006

Acknowledgements

First and foremost, all praise to God who gave me all the help, guidance, and courage to finish my work. May God help me to convey all what I learned for the benefit.

Acknowledgement is due to University of Bridgeport, for providing me with all the support needed to conduct this research.

I am so thankful to my thesis advisor, Dr. Ausif Mahmood, for his continuous help and guidance throughout the course of this thesis and all of my research. As my thesis advisor, he put every possible effort to help me achieve my goals. Working with him is an experience that I can never forget. I am also thankful to Dr. Julius Dichter, Dr. Khaled Elleithy, Professor Gonhsin Liu and Dr. Tarek Sobh for their help.

All my thanks and appreciation go to my parent, in the first place, and then to my family for their cooperation, understanding and patience throughout my academic work. My deep thanks go to my dear friend Fahim Uddin for his great help and support since the beginning of my work and up to the last moment. Thanks also to Mark Yu who gave me chance to work this project.

Dedicated

To My parents

Who helped me a lot to pursue my degree in Computer Science at University of Bridgeport.

Contents

Chapters

INTRODUTION

There are a lot of recommendation tools to help marketing or personalization in e-commerce. This paper will show how people set up e-commerce and what kind of information they expect to sell their products. “Section # 3” gives an overview of e-commerce like what is e-commerce, why people start and what kind of information they expect to sell their products. “Section # 4” describes Data Mining (DM) from definition to classic 2 methods. “Section # 5” discusses about what kind of information we need, how we collect, how we organize for DM. “Section # 6” shows association rule mining and collaborative rule mining. “Section # 7” provides overview of case study. “Section # 8” concludes this paper and discusses further study.

ABSTRACT

This is e-commerce to sell wigs and hairpieces. This company has their own off-line business as well. When I started this project they wanted to build not only web-site but also recommendation system. One more request was that they would like to reflect their own business know-how. Now I will show what classic recommendation system is and how I solved their request.

In off-line business, they are talking with customer and they can guess what they want and what will fit to them and then they can recommend some item to customer. In e-commerce, it is difficult to apply customer's request in real-time. What is the most interactive data in web-site? The answer is web log from web server so most of e-commerce use and analyze that for their system. In this paper, we designed database to transfer from off-line knowledge to on-line recommendation system that can help selling product. This paper concern about Collaborative filter and association rule mining which is popular in this field. In this site, we need special database design to solve their request, to add their own experience and data preparation. This paper shows how data mining can help e-business to improve their customer relationship and make intelligent business strategies.

E-commerce and Recommendation application

Most of people have their own sales long time experience and know-how from their off-line business. Those people want to get more intelligent and useful information which is reflected their own from their solution. What is the solution? That is the one which I want to study, recommendation application using data mining. There are lot of different algorithms in this field. I will provide review of data mining and show example of different cases. Also I will explain what my interesting mining method with particular case is.

----------E-commerce

E-Commerce is the term for electronic business transactions, commerce or Internet trade. E-Commerce or e-business, therefore, refers to the business transactions between companies (B2B) or between companies and their customers (B2C) that are wholly or partially conducted over the Internet or similar public or private computer networks. [21]

In E-commerce, the advantage of location is gone, and businesses must depend more heavily on information advantages.[2]

Acquiring new customers, delighting and retaining existing customers, and predicting buyer behavior will improve the availability of products and services and hence the profits. Thus the end goal of any DM exercise in e-commerce is to improve processes that contribute to delivering value to the end customer.[1]

For a successful e-commerce site, reducing user-perceived latency is the second most important quality after good site-navigation quality. The most successful approach towards reducing user-perceived latency has been the extraction of path traversal patterns from past users' access history to predict future user traversal behavior and to prefect the required resources. [1]

----------Recommendation application

Recommender systems are used by E-commerce sites to suggest products to their customers and to provide consumers with information to help them decide which products to purchase. The products can be recommended based on the top overall sellers on a site, on the demographics of the consumer, or on an analysis of the past

buying behavior of the consumer as a prediction for future buying behavior. The forms of recommendation include suggesting products to the consumer, providing personalized product information, summarizing community opinion, and providing community critiques.[2]

Recommender systems include processes that are conducted largely by hand, such as manually creating cross-sell lists, and actions that are performed largely by computer, such as collaborative filtering. Recommender systems enhance E-commerce sales in three ways:

Converting Browsers into Buyers: Visitors to a Web site often look over the site without purchasing anything. Recommender systems can help consumers find products they wish to purchase.

Increasing Cross-sell: Recommender systems improve cross-sell by suggesting additional products for the customer to purchase. If the recommendations are good, the average order size should increase. For instance, a site might recommend additional products in the checkout process, based on those products already in the shopping cart.

Recommender systems are a technology that helps merchandisers implement a one-to-one marketing strategy. The recommender system analyzes a database of consumer preferences to overcome the limitations of segment-based mass marketing by presenting each customer with a personal set of recommendations. Of course, recommender systems are not a complete solution. It is still necessary to record and use other consumer data, such as preferred credit card and shipping address, to deliver complete one-to-one service to consumers.[2]

Recommender systems are technologies that can help businesses decide to whom to make an offer. Such systems allow search engines and advertising companies to suggest advertisements or offers to display based on consumer behavior.[2]

Online recommendations are preferred because they can respond immediately to the consumer's preferences. Most of the recommender processes mentioned above can be performed entirely online.[2]

Cross-Selling: Product-Associated Recommendations Suggestive selling is particularly effective when the seller knows the current interests of the buyer. Retailers arrange products to enhance cross-selling by placing complementary items in close proximity. On-line retailers are freed from physical layout and can directly suggest products related to the one a customer is viewing. By using the targeted customer's current navigation as an ephemeral indication of interest, such systems use item-to-item correlation and community purchase history to passively display suggestions to the targeted customer. Many different recommender applications use the context of a

current product or several current products to recommend other products using a variety of recommender methods. This popularity is partly due to the variety of inputs that can be used to generate such recommendations, including anonymous purchase histories, customer purchase histories, ratings, product attributes and expert opinions. Product-associated recommendations are particularly well suited for passive delivery since they can be integrated into a product information page.

Building Long-Term Relationships: Deep Personalization The goal of most retail businesses is to develop long-term relationships with customers that lead to higher lifetime values and greater competitive barriers. Deep personalization, based on a customer's history of preferences, purchases, or navigation, is the strongest and most difficult type of personalization to implement. Deep personalization is common already in web advertising and is becoming more widely used in e-commerce now that collaborative filtering recommendation engines are readily available. By utilizing collaborative filtering's ability to match the targeted customer's history with histories of other customers, deep personalization is able to generate persistent, personalized suggestions or predictions. Deep personalization builds a customer relationship over time, leveraging the history developed to provide increasingly better recommendations. Unlike notification services that require manual updating, deep personalization updates the user profile whenever the customer interacts with the merchant. Deep personalization systems can use user-to-user correlation, attribute-based systems with a learning module to identify user interests, or a combination of the two.[2]

----------Consumer-Centered Recommendations

One important difference between a recommender system and a traditional data mining system is that the end-user for the recommender system is the consumer. This difference leads to several desirable properties for recommender system algorithms that have not been explored in data mining algorithms. Some recommendations are most valuable when they apply to a group of consumers rather than an individual. For instance, movies are most often attended socially. The choice of the right "date movie" can be very important. Recommenders can be used to select products that maximize the value of the product to a group of people. Some systems already support simple versions of this idea, such as selecting a movie that two people will like. Future multi-user recommendation systems will let customers control how the recommender system balances their interests in choosing a product. Should it choose a movie that neither will hate? That she will love? That they will

both like? Different algorithms are needed for each of these scenarios to maximize the social good for the group, according to their needs. The success of a recommender system should be measured by how effectively the system helps customers make decisions that they, retrospectively, consider correct. One interface innovation that can help customers decide when to follow recommendations is to explain the recommendation to the customer, just as many machine-learning algorithms explain their results to their users. Researchers are currently experimenting with several different explanation models, including summarizing the data behind recommendations, explaining recommendations indirectly (e.g., indicating which of the customer's own preferences most strongly led to a particular recommendation), and providing "persuasive" evidence about the system's success in similar recommendations. Simpler explanation systems display a brief capsule of the amount of data or expected variance in a prediction. Additional research is needed regarding explanation algorithms for other recommender algorithms, and on the effectiveness of explanations in helping customers decide among recommendations and increasing customer confidence in using recommender systems.

----------Connecting Recommenders to Marketers

Recommender systems are currently used as virtual salespeople, rather than as marketing tools. The recommender system makes good decisions locally, but it is neither able to give the type of feedback needed for marketing professionals, nor is it able to accept input from marketing professionals about global preferences. In the future, recommender systems must integrate marketing features more completely. Systems based on association rules or other non-personal data mining approaches build models of customer preference and must then filter out products the customer already owns to avoid seemingly trivial recommendations. Personalized recommender systems include customer ratings and purchases in the process and thereby avoid this challenge, but they too could benefit from more effective filtering of recommendations. Frequently either the marketer or the customer wants a recommendation set that is restricted based on specific criteria. For instance, a consumer might want recommendations only for products that are currently in stock, or a marketer might want to avoid recommending an R-rated movie to an eleven year old. Today's recommender applications must post filter recommendations or pre-filter candidates for recommendation. Research is needed to better integrate the recommendation process with other forms of product filtering to allow more

integrated solutions.

Recommender systems should also be integrated with the marketer's reporting systems. Recommender systems target each individual customer differently, making it difficult to produce the reports to which marketing professionals are accustomed. These reports usually partition the population into a manageable number of segments. One way to bring these two worlds together would be to use the people-to-people correlations implemented by some recommender system algorithms to create segments for the reports. Open questions include "How can names be assigned to the automatically generated segments?" and "Are automatically generated segments more useful for managing marketing campaigns than traditional segments?"

Recommender systems can be made more useful as marketing systems in other ways as well. Current recommender systems are mainly "buy-side" systems; that is, they are designed to work on behalf of the consumer to help him decide what products he should purchase. However, modern marketing is designed not only to maximize utility to the customer but also to maximize value to the business at the same time. The recommender system could produce an indication of the price sensitivity of the customer for a given product. For instance, one customer might be willing to purchase the product at a price that would earn the site ten cents of profit while another customer might purchase the same product at a one-dollar profit. Furthermore, this price sensitivity indication could be combined with a lifetime value model (e.g., Mani, 1999) to determine when it is in the best interest of the site to sell a product at a lower price–potentially even below cost–to increase the customer's loyalty and thereby her long term value to the site. There are challenging ethical and practical issues in implementing systems like these that use information from studying the customer to implement price discrimination. The ethics of opinion ownership may lead sites to directly compensate customers for their information (Avery, 1999).

How will recommender systems that are serving marketers evolve to protect the trustworthiness that makes them valuable to customers? Recommender systems emerged as agents to help customers find products to purchase. These recommenders were implemented by sites to create additional value for their customers, so those customers would be more loyal to the site. Now, some recommenders are evolving towards marketing systems that help sites sell high profit products. One way to counter this evolution is to create verifiably unbiased recommender systems that consumers trust to recommend on their behalf. In the mid 1990s Firefly proposed creating such a branded recommender system.

Consumers may be suspicious that an unscrupulous Web site might modify the outputs of such a recommender before displaying them. How can a trusted recommender validate itself to consumers through a Web client? Is it possible to build portable recommenders that can run on consumer's palmtop computing devices, serving only the consumer?

For small collections of items, such as movies, it is possible to produce recommendations for a large set of items in advance and simply display them on-site. In theory, an off-line recommender could store the ratings profiles from a neighborhood of customers, but in practice the level of ratings sparsity makes this approach infeasible. Is it possible to build an entire off-line recommender system that can accept new ratings from the target customer and generate real-time recommendations? Such a recommender might include the use of filtering agents trained to replicate the rating pattern and the generation of compressed ratings datasets through dimensionality reduction techniques. Consumers would know they could trust their own live recommender system loaded with a highly compact representation of the opinions of a community of trusted consumers.

Recommender systems allow businesses to leverage their customer history to create more personalized experiences for their customers. Those customers will quickly discover that the business that "knows them best" is the one that can serve them most effectively, recommending the right products rather than treating them like strangers. [2]

To construct a recommender system, two information filtering methods for providing the recommended information are considered: (1) by analyzing the information content, i.e., content-based filtering, and (2) by referencing other users' access behaviors, i.e., collaborative filtering. However, most of the existing approaches consider only one of the filtering techniques in their recommending processes. In order to recommend the information to a particular user, the original collaborative filtering technique relies on the Nearest Neighbor (NN) clustering algorithm to compare and search for other users' ratings which are

closely resemble to the user's profile. [4]

Recommender systems are personalized information filtering technology used to either predict whether a particular user will like a particular item (prediction problem) or to identify a set of N items that will be of interest to a certain user (top-N recommendation problem).[6]

Over the years, various approaches for building recommender systems have been developed that utilize either demographic, content, or historical information.[6] Recommender systems combine ideas from artificial intelligence, machine learning,

information retrieval and user profiling and have been applied successfully in a variety of domains including e-commerce and online shopping.[7]

The recommendation engine takes a collection of frequent item sets as input and generates a recommendation set for a user by matching the current user's activity against the discovered patterns. The recommendation engine is an online process, therefore its efficiency and scalability are of paramount importance. [9]

The recommendation value of each candidate page view is based on the confidence of the corresponding association rule whose consequent is the singleton containing the page view to be recommended. If the rule satisfies a specified confidence threshold requirement, then the candidate page view is added to the recommendation set.[9]

----------What is Data Mining

Data mining (DM) has as its dominant goal, the generation of non-obvious yet useful information for decision makers from very large databases. [1]

The term *data mining* is used to describe the collection of analysis techniques used to infer rules from or build models from large data sets. One of the best-known examples of data mining in commerce is the discovery of association rules – relationships between items that indicate a relationship between the purchase of one item and the purchase of another.[2]

Data mining or knowledge discovery in databases (KDD) has emerged recently as an active research area for extracting implicit, previously unknown, and potentially useful information from large databases.[4]

Given a truly massive amount of data, the challenge in data mining is to unearth hidden relationships among various attributes of data and between several snapshots of data over a period of time. These hidden patterns have enormous potential in predictions and personalizations in e-commerce. Data mining has been pursued as a research topic by at least two communities: the statisticians and neural network. We now present a brief overview of some of the features of each of these approaches.[1]

----------A review of data-mining methods

----------statistics

By strict definition "statistics" or statistical techniques are not data mining. They were being used long before the term data mining was coined to apply to business applications. However, statistical techniques are driven by the data and are used to discover patterns and build predictive models.[22]

Statisticians consider the causal relationship between the dependent variables and independent variables as proposed by the user (usually the domain expert), and endeavor to capture the degree and nature of dependence between the variables.

Knowing statistics in your everyday life will help the average business person make better decisions by allowing them to figure out risk and uncertainty when all the facts either aren't known or can't be collected. Even with all the data stored in the largest

of data warehouses business decisions still just become more informed guesses. The more and better data and the better understanding of statistics make the better the decision that can be made. Today data mining has been defined independently of statistics though “mining data” for patterns and predictions is really what statistics is all about.

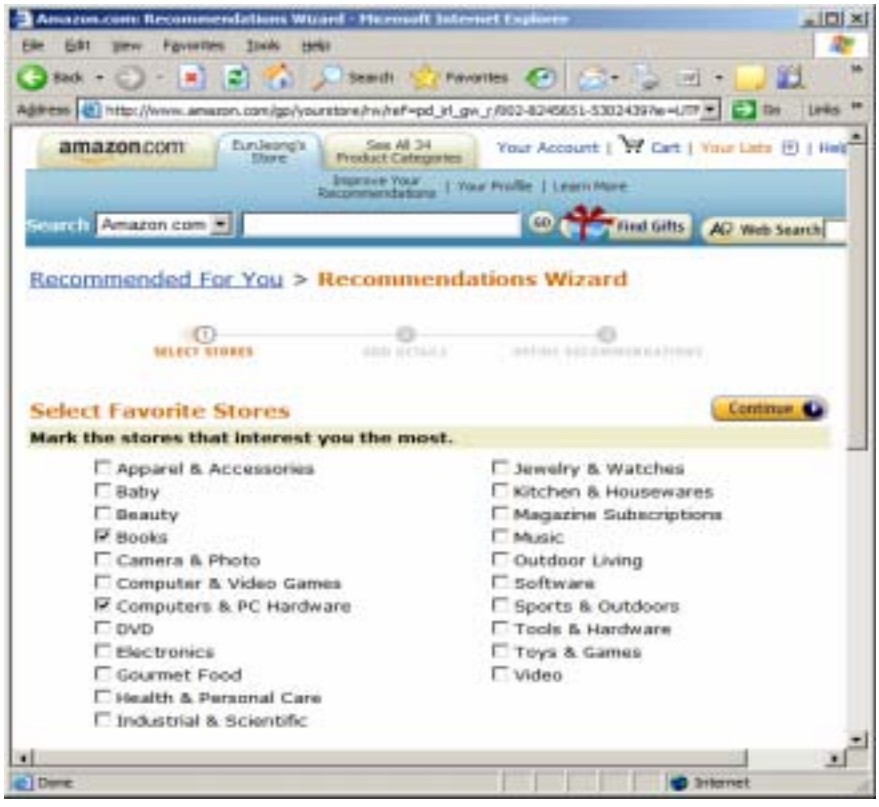
----------Neural Network

Neural networks are predominantly used to learn linear and nonlinear relationships between variables of interest. The architecture, in general, consists of a perceptron with input and output nodes with weighted edges connecting the two nodes.[1] Neural networks do have disadvantages that can be limiting in their ease of use and ease of deployment, but they do also have some significant advantages. Foremost among these advantages is their highly accurate predictive models that can be applied across a large number of different types of problems. To be more precise with the term “neural network” one might better speak of an “artificial neural network”. True neural networks are biological systems that detect patterns, make predictions and learn. The artificial ones are computer programs implementing sophisticated pattern detection and machine learning algorithms on a computer to build predictive models from large historical databases.[10]
Neural networks are very powerful predictive modeling techniques but some of the power comes at the expense of ease of use and ease of deployment. Neural networks create very complex models that are almost always impossible to fully understand even by experts. The model itself is represented by numeric values in a complex calculation that requires all of the predictor values to be in the form of a number. The output of the neural network is also numeric and needs to be translated if the actual prediction value is categorical. They have been used in all facets of business from detecting the fraudulent use of credit cards and credit risk prediction to increasing the hit rate of targeted mailings. [10]

This initial step involves the collection of data sets for executing the data mining algorithms. This step includes the data reduction and selection techniques to improve the efficiency of the data mining algorithms.[4] After the data is collected, it must be prepared, cleaned, and stored in a data warehouse.

----------Manually selected data

Applications use various kinds of data about individual customers. Many applications classify the data into two basic types: factual—who the customer is - and transactional - what the customer does.[23]
Now I will present factual data. The factual data includes demographic information such as name, gender, birth date, address, salary, and social security number. A key issue in developing recommendation applications is constructing accurate and comprehensive customer profiles based on the collected data. Some e-commerce sites include customer's preference like amazon.com. This basic information is used to use as first filter to get recommendation information.



----------Web Server logs

The information available in log files is often used to determine what profiling can be dynamically processed in the background and indexed into the dynamic generation of HTML, and what performance can be expected from the servers and network to support customer service and make e-business interaction productive.[1]

We can use the data for pattern discovery techniques such as clustering, association rule-mining, and sequential pattern discovery. The log data are collected automatically by the Web and application servers represent the fine-grained navigational behavior of visitors. [1]

The Web server is implemented on the HTTP specification and has the following functions: listening for HTTP requests on a network, receiving HTTP requests made by user agents (usually Web browsers), serving the requests (accessing the database), and returning HTTP response that contains the requested resources. [4]

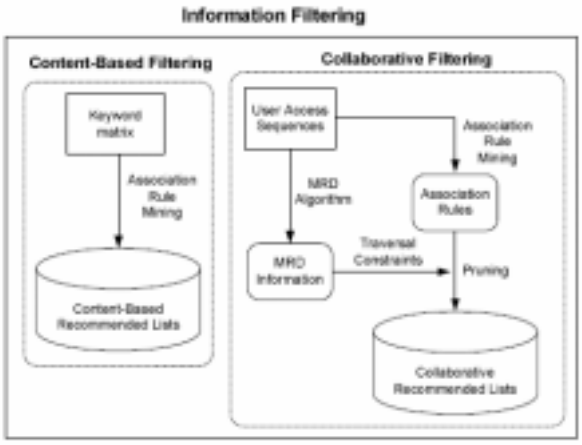
Each time a page in your web site is requested, your web server drops a line on the access log. The access log is basically a relation (or table, if you like) with a few columns that record what is regarded as important information. This data can then be mined for different purposes.[5]

The corresponding data mining problem is *"Based on the sites web log, build a recommendation model that, given a set of visited pages, indicates a list of interesting links to the user".[5]*

Many of the technologies used in the actual recommender systems studied are fairly simple database queries. Automatic recommender systems, however, use a wide range of techniques, ranging from nearest neighbor algorithms to Bayesian analysis.[2]

Two types of information filtering can be accomplished by using data mining: content-based filtering and collaborative filtering. In the proposed framework, the association rule mining technique is applied as the content-based filtering where the data set is the keyword matrix. Collaborative or social-based filtering retrieves the information for a particular user by referring to other user evaluations on the information content.[4]

Information Filtering via Data Mining: This step is the core process of the recommender system framework, where the data sets are analyzed and the data mining algorithms are applied as the information filtering tools to generate and discover any useful and interesting recommended outputs.[4]



----------Collaborative filtering

The earliest recommenders used nearest-neighbor collaborative filtering algorithms (Resnick et al. 1994, Shardanand et al. 1995). Nearest neighbor algorithms are based on computing the distance between consumers based on their preference history. Predictions of how much a consumer will like a product are computed by taking the weighted average of the opinions of a set of nearest neighbors for that product.[2]

Clustering techniques work by identifying groups of consumers who appear to have similar preferences. Once the clusters are created, predictions for an individual can be made by averaging the opinions of the other consumers in that cluster. Clustering techniques can also be applied as a “first step” for shrinking the candidate set in a nearest neighbor algorithm or for distributing nearest-neighbor computation across several recommender engines.[2]

Recommender systems using user-to-user correlation recommend products to a customer based on the correlation between that customer and other customers who have purchased products from the E-commerce site. This technology is often called “collaborative filtering” because it originated as an information filtering technique that used group opinions to recommend information items to individuals (Resnick et al., 1994; Hill et al., 1995; Shardanand and Maes, 1995; Konstan et al., 1997). We use the word correlation in the name of this technique, thus hinting at nearest-neighbor techniques based on linear correlation, the technique can be implemented with many other technologies as well (Breese et al., 1998).[2]

Collaborative filtering (CF), which relies on historical information, is probably the most successful and widely used technique for building recommender systems. Referred to as user-based, relies on the fact that each person belongs in a larger group of similarly behaving individuals.[6]

User-based collaborative filtering is the most successful technology for building recommender systems to date and is extensively used in many commercial recommender systems. In general, user-based systems compute the top-N recommended items for a particular user by following a three-step approach. In the first step, they identify the k users in the database that are the most similar to the active user. During the second step, they compute the union of the items purchased by these users and associate a weight with each item based on its importance in the set. Finally, in the third step, they select the N items from that union that have the highest weight and have not already been purchased by the active user as the items to be recommended.[6]

Collaborative filtering algorithms make use of simple ratings-based profiles and generate recommendations on the basis of similarity between the target user and neighboring profiles.[7]

----------Association Rule mining

Association rules have been used for many years in merchandising, both to analyze patterns of preference across products, and to recommend products to consumers

based on other products they have selected. An association rule expresses the relationship that one product is often purchased along with other products. The number of possible association rules grows exponentially with the number of products in a rule, but constraints on confidence and support, combined with algorithms that build association rules with item sets of n items from rules with n-1 item sets, reduce the effective search space. [2]

A transaction is a set of items. An association rule is a rule of the form X => Y, where X and Y are sets of items; X and Y are called respectively the body and the head of the rule. The intended meaning of this rule is that the presence of (all of the items of) X in a transaction implies the presence of (all of the items of) Y in the same transaction with some probability. Each association rule has two measures relative to a given set of transactions: its confidence and its support. Confidence is the percentage of transactions that contain Y among transactions that contain X; support is the percentage of transactions that contain both X and Y among all transactions in the input data set. In other words, the confidence of a rule measures the degree of the correlation between item sets, while the support of a rule measures the significance of the correlation between item sets. The traditional association rule mining problem definition is: given a set of transactions, a user-specified minimum support and minimum confidence, find all association rules that are above the user-specified minimum support and minimum confidence. This problem definition is well suited to uncovering aggregate user behaviors as required in traditional market basket analysis. [3]

When mining association rules there are mainly two problems to deal with: First of all there is the algorithmic complexity. The number of rules grows exponentially with the number of items. Fortunately today's algorithms are able on minimal thresholds for quality measures on rules. Second, interesting rules must be picked from the set of generated rules. This might be quite costly because the generated rule sets normally are quite large – e.g. more than 100,000 rules are not uncommon –and in contrast the percentage of useful rules is typically only a very small fraction.[8]

Association rules can form a very compact representation of preference data that may improve efficiency of storage as well as performance. They are more commonly used for larger populations rather than for individual consumers, and they, like other learning methods that first build and then apply models, are less suitable for applications where knowledge of preferences changes rapidly. Association rules have been particularly successfully in broad applications such as shelf layout in retail stores. By contrast, recommender systems based on nearest neighbor techniques are easier to implement for personal recommendation in a domain where consumer

opinions are frequently added, such as on-line retail.[2]

Other applications use item-to-item correlation to identify items frequently found in “association” with items in which a customer has expressed interest. Association may be based on co-purchase data, preference by common customers, or other measures. In its simplest implementation, item-to-item correlation can be used to identify "matching items" for a single item, such as other clothing items that are commonly purchased with a pair of pants. More powerful systems match an entire set of items, such as those in a customer's shopping cart, to identify appropriate items to recommend. Item-to-item correlation recommender applications usually use current purchases or other current interests rather than long-term customer history, which make them particularly well-suited for recommending gifts. A customer merely needs to identify some other items liked by the recipient to elicit gift recommendations tailored to the recipient rather than the giver.[2]

Product-associated recommendations allow businesses to respond to each customer's current interests and allow the natural associations among different products to guide customers to the right purchase. These recommendations combine the helpfulness of a knowledgeable salesperson who can recommend items to match ones of interest with the layout of a good store where complementary items are conveniently shelved near each other, even if that means they are shelved in many locations. [2]

Applying the association rule mining for mining user access patterns allows the prediction of the Web pages to include additional non-consecutive Web pages, and thus enhances the prediction performance in terms of precision and recall. Applying data mining algorithms, e.g., association rule mining, for the information filtering techniques provides efficiency, since the recommended list of information can be generated prior to the recommendation process. Therefore, the response time of the recommender system could be improved.[4]

Content-based filtering is achieved by applying the association rule mining technique on the keyword matrix. The results are a set of rules in the form of “IF (preconditioned set of Web pages) THEN (post-conditioned Web page).” The content-based filtering rules are such that the preconditioned Web pages imply the post conditioned Web page based on the similarity in the keywords.[4]

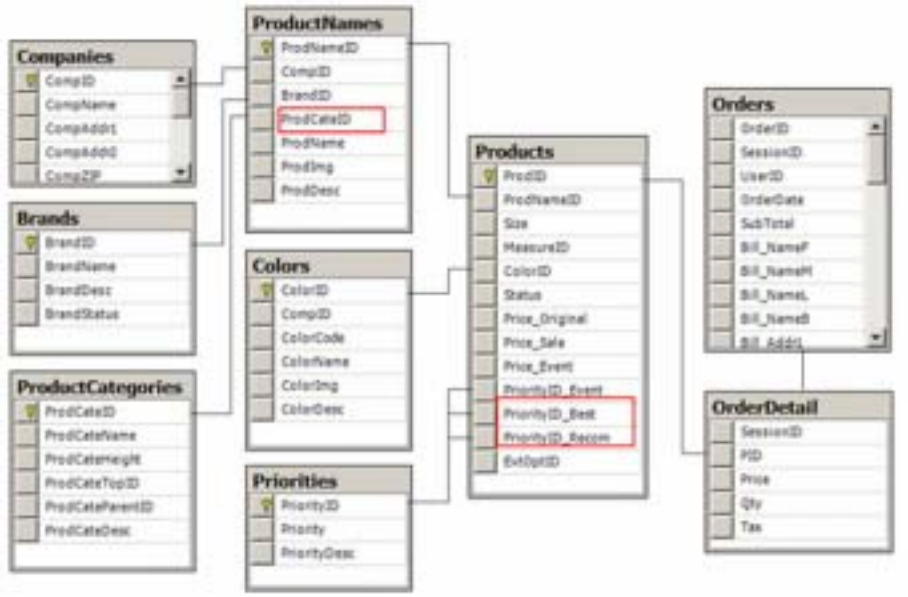
Association rules capture the relationships among items based on their patterns of co-occurrence across transactions. In the case of Web transactions, association rules capture relationships among page views based on the navigational patterns of users. Association rules which satisfy a minimum *confidence* threshold are then generated from the frequent item sets. A problem with using a global minimum support

threshold is that the discovered patterns will not include “rare” but important items which may not occur frequently in the transaction data. This is particularly important in the current context: when dealing with Web usage data, it is often the case that references to deeper content or product oriented pages occur far less frequently that those of top level navigation-oriented pages.[9]

In the following section we present 3 E-commerce businesses that use one or more variations of recommender system technology in their web sites. For each site, and each variation, we give a brief description of the features of the system.

----------Beautysassy.com

This is e-commerce to sell wigs and hairpieces. This company has their own off-line business as well. When I started this project they wanted to build not only web-site but also recommendation system. One more request was they would like to reflect their own business know-how. Now I will show what classic recommendation system is and how I solved their request.



Typically, most of recommendation system use data of Orders and OrderDetail(This is purchase history of users) tables for data mining but start point any system doesn't have any history so that I designed database to collect data from marketers. This figure is database design for this web-site. You can see ProductNames.ProdCateID, Products.PriorityID_Best and Products.PriorityID_Recom. These three fields are important in this system. I try to make narrow data range using "Products.PriorityID_Recom" field.

Priorities table contain page category level for recommend items where can place in.

Table - dbo.Priorities

PriorityID	Priority	PriorityDesc
1	home	Home Page
2	category	NULL
3	product	NULL
4	NONE	NONE
NULL	NULL	NULL

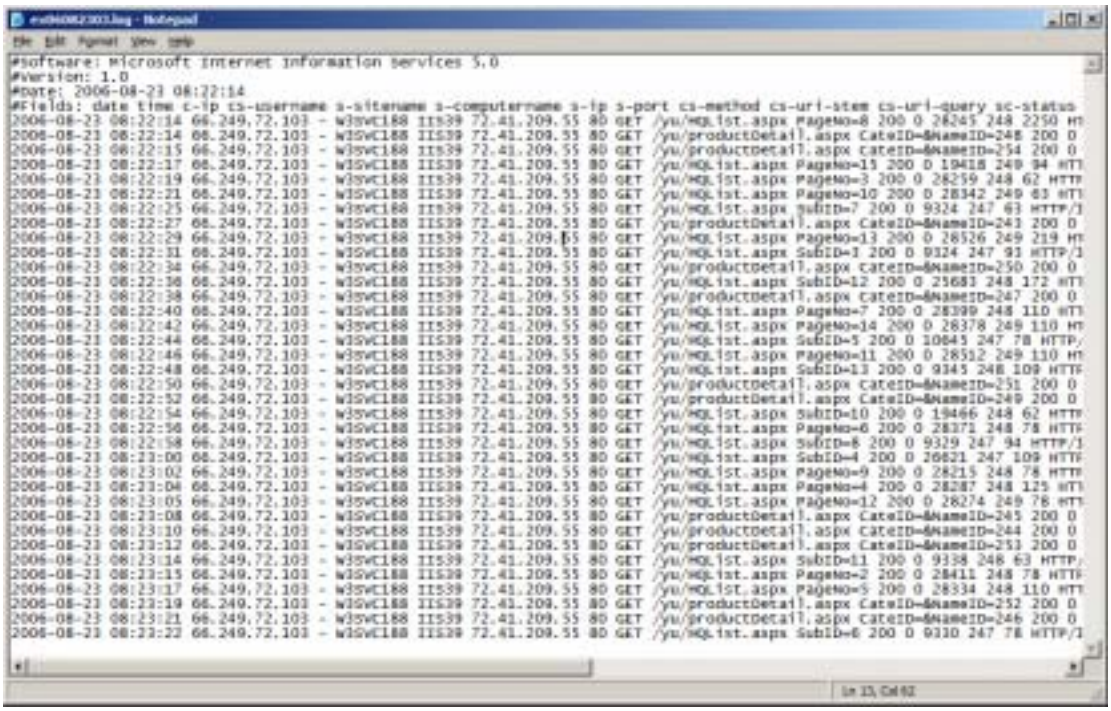
In Products table, I set priority of page for each item from marketers and I can get recommendation items in their opinion. So I can use two different selected data, one is from marketer in other words, it reflects their own experience, another is from users that will be users' purchase history (Table : Order and OrderDetail). If "PriorityID_recom" is "category" it cannot show in "producty"page.

This is how I make data range narrow.

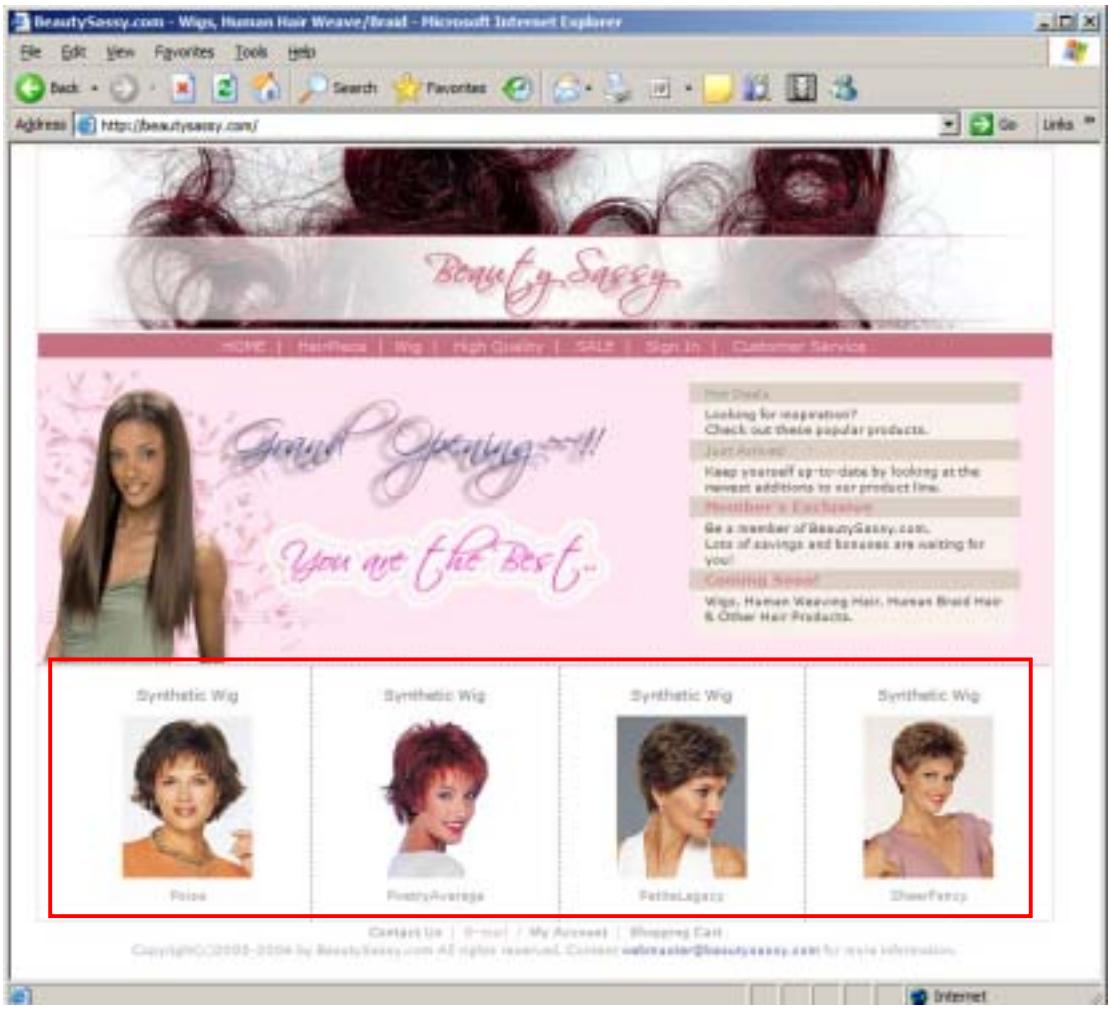
Table - dbo.Products

...rsID	ColorID	Status	Price_Original	Price_Sale	Price_Event	PriorityID_Event	PriorityID_Best	PriorityID_Recom
	1	T	NULL	14.99	NULL	4	1	1
	2	T	NULL	14.99	NULL	4	3	1
	7	T	NULL	14.99	NULL	4	4	2
	8	T	NULL	14.99	NULL	4	4	4
	9	T	NULL	14.99	NULL	1	1	1
	4	T	NULL	14.99	NULL	4	4	4
	1	T	NULL	17.99	NULL	4	4	4
	2	T	NULL	17.99	NULL	1	1	1
	7	T	NULL	17.99	NULL	4	4	4
	8	T	NULL	17.99	NULL	4	4	1
	1	T	NULL	11.99	NULL	4	1	1
	2	T	NULL	11.99	NULL	1	3	4
	3	T	NULL	11.99	NULL	4	3	4
	7	T	NULL	11.99	NULL	1	1	1
	8	T	NULL	11.99	NULL	4	4	4
	9	T	NULL	11.99	NULL	4	4	4
	4	T	NULL	11.99	NULL	1	2	4
	1	T	NULL	47.99	NULL	1	2	4
	2	T	NULL	47.99	NULL	4	2	4
	3	T	NULL	47.99	NULL	4	2	4
	7	T	NULL	47.99	NULL	1	4	3
	4	T	NULL	47.99	NULL	4	4	3
	83	T	NULL	52.99	NULL	4	4	3
	1	T	NULL	54.99	NULL	4	4	3
	2	T	NULL	54.99	NULL	4	4	3
	3	T	NULL	54.99	NULL	4	4	4
	7	T	NULL	54.99	NULL	4	4	4

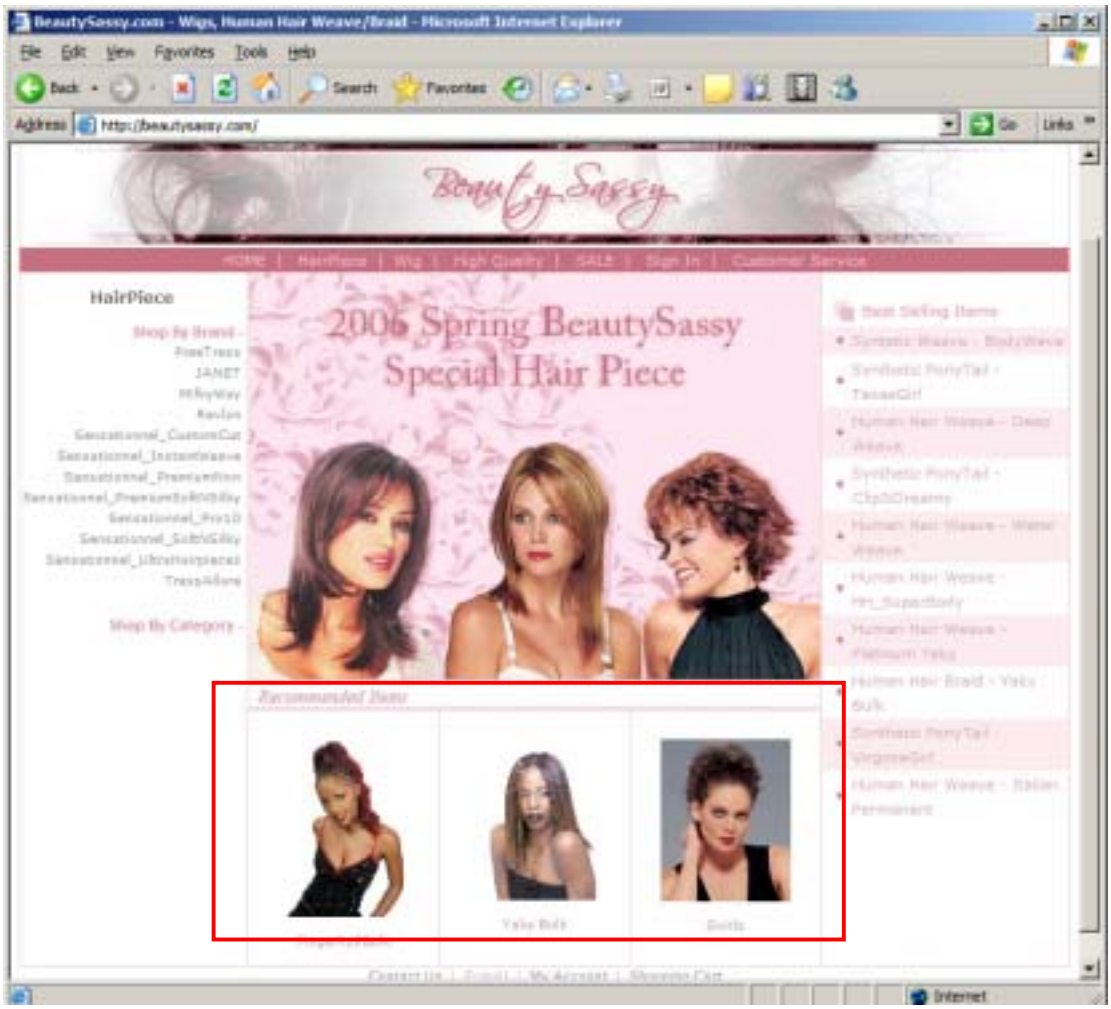
This is web log which write from web server. We can get IP address, username, URI and references. First, we have to clean this web log before we use for data mining. We need to implement some program for data cleansing and we can save information in our database. I can get from this file I can discovery pattern of user, which category selected often and which product people are interested in because I use "get" method which is including whole URI.



I have three different pages for recommendation items. Specially, I am using collaborative filter algorithm in this web site with some special field which I already explained.



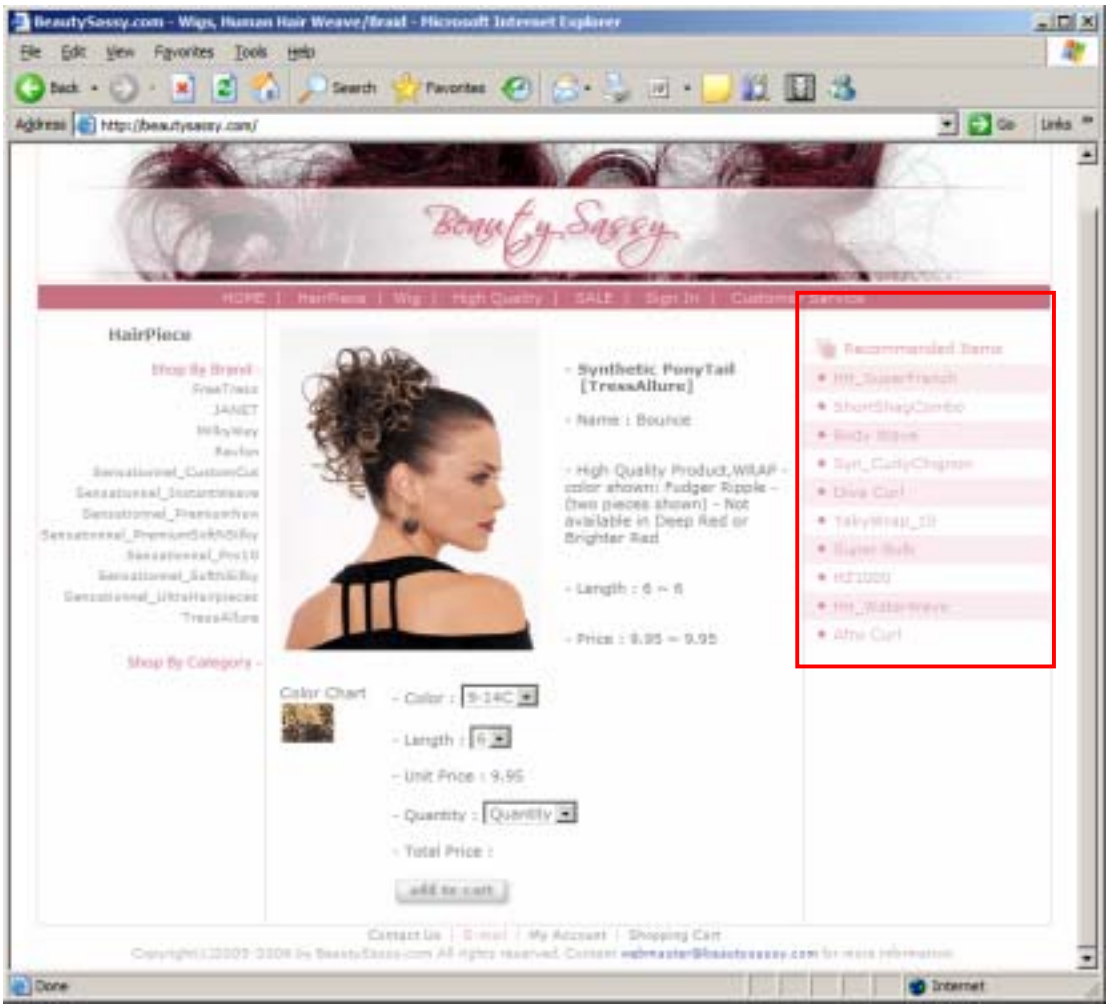
I got 4 items in “home” which includes category and product in priorityID_recom field. It doesn’t care about product categories, for example wig or hairpieces but next page is different.



When you select hairpieces on top main menu web server record this line (below box) in web log. We can see which product category select.

2006-08-23 06:23:01 61.74.213.55 - W3SVC188 IIS39 72.41.209.55 80 GET /yu/category.aspx CateID=1 200 0 24526 487 1266 HTTP/1.1
Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.0)
ASP.NET_SessionId=ycsxjd45vihd5j55aesz2z55
http://www.beautysassy.com/yu/home.aspx

This recommendation items are belong to hairpieces (product category) and category page (page category or product detail: in Priority table). These conditions make data range narrow.



This is most detail page in this e-commerce. The condition of this page is " product category = hairpieces" and "page category = product detail".

2006-08-23 08:22:50 66.249.72.103 - W3SVC188 IIS39 72.41.209.55 80 GET /yu/productDetail.aspx CateID=1&NameID=329 200 0 19517 265 187 HTTP/1.1 Mozilla/5.0+(compatible;+Googlebot/2.1;++http://www.google.com/bot.html) - -

Further more, these recommendation items are in product categories of users' favorite which are from web log.

----------Amazon.com [2]

We will focus here on recommender systems in the *book* section of Amazon.com. Customers Who Bought: Like many E-commerce sites, Amazon.com™ (www.amazon.com) is structured with an information page for each book, giving details of the text and purchase information. The Customers Who Bought feature is found on the information page for each book in their catalog. It is in fact two

separate recommendation lists. The first recommends books frequently purchased by customers who purchased the selected book. The second recommends authors whose books are frequently purchased by customers who purchased works by the author of the selected book. Your Recommendations: Amazon also encourages direct feedback from customers about books they have read. Customers rate books they have read on a 5-point scale from “hated it” to “loved it.” After rating a sample of books, customers may request recommendations for books that they might like. At that point, a half dozen non-rated texts are presented that correlate with the user’s indicated tastes. Figure 1 shows a sample screen from your recommendations.

Eyes: The Eyes feature allows customers to be notified via email of new items that have been added to the Amazon.com catalog. Customers enter requests based upon author, title, subject, ISBN, or publication date information. Customers can use both simple and more complex Boolean-based criteria (AND/OR) for notification queries. One of the interesting variations of the Eyes system allows requests to be directly entered from any search results screen, creating a persistent request based on the search. Amazon.com Delivers: Amazon.com Delivers is a variation on the Eyes feature. Customers select checkboxes to choose from a list of specific categories/genres (Oprah books, biographies, cooking). Periodically the editors at Amazon.com send their latest recommendations by email to subscribers in each category. Bookstore Gift Ideas: The Gift Ideas feature allows customers to receive recommendations from editors. Customers pick a category of books for which they would like some suggestions. By navigating to that section of the “Gift Department,” they can view a general list of recommendations created by the editors of Amazon.com. They also can select to view recommendations in one of a predefined list of categories including Globetrotter, Entrepreneur, and Teens. In many ways this serves as an online version of the Amazon.com Delivers feature discussed earlier. However, customers can be provided with recommendations anonymously since there is no need to register with the site as there is with Delivers. Customer Comments: The Customer Comments feature allows customers to receive text recommendations based on the opinions of other customers. Located on the information page for each book is a list of 1-5 star ratings and written comments provided by customers who have read the book in question and submitted a review. Customers have the option of incorporating these recommendations into their purchase decision. Furthermore, customers can “rate the comments.” With each comment is the question “Did this comment help you.” Customers may indicate yes or no. Results are tabulated and reported such as “5 of 7 people found the following review helpful.” Purchase Circles: The Purchase Circles feature allows customers to

view the "top 10" list for a given geographic region, company, educational institution, government or other organization. For example, a customer could request to see what books are the best sellers for customers at Oracle, MIT, or residents of New York City. Purchase Circles provide another “fellow customer” form of recommendations by allowing customers not only to see what others are reading but also to personalize the recommendations by allowing them to select a “domain” with which they associate themselves. Customers can view Purchase Circles by navigating to the Circle that interests them.

----------eBay.com[2]

Feedback Profile: The Feedback Profile feature at eBay.com™ (www.ebay.com) allows both buyers and sellers to contribute to feedback profiles of other customers with whom they have done business. The feedback consists of a satisfaction rating (satisfied/neutral/dissatisfied) as well as a specific comment about the other customer. Feedback is used to provide a recommender system for purchasers, who are able to view the profile of sellers. This profile consists of a table of the number of each rating in the past 7 days, past month, and past 6 months, as well as an overall summary (e.g., 867 positives from 776 unique customers). Upon further request, customers can browse the individual ratings and comments for the sellers.

Personal Shopper: The personal shopper feature of eBay allows customers to indicate items they are interested in purchasing. Customers input a “short term” (30/60/90 days) and search on a set of keywords of their choosing, including their price limit. On a periodic basis (one or three day intervals) the site performs the customer’s search over all auctions at the site and sends the customer an email with the results of this search.

Application Model	Functional I/O			Recommendation Method	Design Issues	
	Targeted Customer Input	Community Input	Output		Degree of Personalization	Delivery
Broad Rec. List						
Amazon Bookstore Gift Ideas	Explicit navigation	Item attributes External item popularity	Suggestion	Manual selection	Non-personalized	Pulling unordered list and expert narrative
Comments and Ratings						
Amazon Customer Comments	Implicit navigation	Ratings Text comments	Prediction Ratings Reviews	Statistical summarization	Non-personalized	Passive delivery of comments, individual ratings of other customers, and a predicted rating.
eBay Feedback Profile	Implicit navigation	Ratings Text comments	Prediction Ratings Reviews	Statistical summarization	Non-personalized	Passive delivery of comments, individual ratings of other customers, and a predicted rating.
Notification Service						
Amazon Eyes	Keyword/Attribute	Item attributes	Suggestion	Attribute based	Persistent	Pushing single recommendation
Amazon.com Delivers	Keyword/Attribute	Item attributes	Reviews	Manual selection	Persistent	Pushing expert narrative
eBay Personal Shopper	Keyword/Attribute	Item attributes	Suggestion	Attribute based	Persistent	Pushing unordered list
Product-Associated						
Amazon Customers who Bought	Implicit navigation	Purchase history	Suggestion	Item-to-Item correlation	Ephemeral	Passive delivery of unordered list
Deep Personalization						
Amazon Your Recommendations	Purchase History Ratings	Ratings Purchase history	Prediction Suggestion	User-to-User correlation	Persistent	Pulling (un)ordered list

Conclusion and further work

Successful e-business needs intelligent recommendation system. This paper describe what is classic data mining method and what is the popular algorithm in this field and what kind of data system needs and how system can collect data for data mining. System is working with their data still people who operate this system want to reveal their opinion manually. This recommendation system in this website designed automation operation as well as manual operation as their request. This recommendation system helps users find items they want to buy and is available cross-selling in product detail page. This system is going to be improved mining technique with more various data such as web log, selected items and purchase history and searched optimized recommended items.

References

[1] N R SRINIVASA RAGHAVAN “Data mining in e-commerce: A survey” Department of Management Studies, Indian Institute of Science. (April/June 2005, pp. 275–289.)

[2] J. Ben Schafer, Joseph A. Konstan, and John Riedl “E-Commerce Recommendation Applications” Data Mining and Knowledge (January-April 2001)

[3] Weiyang Lin, Sergio A. Alvarez and Carolina Ruiz “Collaborative Recommendation via Adaptive Association Rule Mining” (2000)

[4] A Data Mining Framework for Building A Web-Page Recommender System

[5] Jorge, A., Alves, M. A. and Azevedo, P. “Recommendation with Association Rules: A Web Mining Application” in Proceedings of Data Mining and Werehouses, a sub-conference of information society 2002, EDS. Mladenic, D., Grobelnik, M., Josef Stefan Institute. (October 2002)

[6] Eui-Hong (Sam) Han and George Karypis “Feature-Based Recommendation System” Conference on Information and Knowledge Management (2005)

[7] Barry Smyth, Kevin McCarthy, James Reilly, Derry O'Sullivan, Lorraine McGinty and David C. Wilson “Case-Studies in Association Rule Mining for Recommender Systems” (2005)

[8] Jochen Hipp, Ulrich Güntzer and Gholamreza Nakhaeizadeh “Algorithms for Association Rule Mining – A General Survey and Comparison” SIGKDD Explorations (2000)

[9] Bamshad Mobasher, Honghua Dai, Tao Luo and Miki Nakagawa “Effective Personalization Based on Association Rule Discovery from Web Usage Data” Web Information and Data Management (2001)

[10] Why and What Is Data Mining?

[11] Jaideep Srivastava, Robert Cooley, Mukund Deshpande and Pang-Ning Tan “Web

Usage Mining: Discovery and Applications of Usage Patterns from Web Data" SIGKDD Explorations

[12] Beautysasy.com

[13] Amazon.com

[14] eBbay.com

[15] M. Deshpande and G. Karypis. "Item-Based Top-*N* Recommendation Algorithms" ACM Trans. Infomation System. (2004)

[16] Mei-Ling Shyu, Choochart Haruechaiyasak, Shu-Ching Chen and Na Zhao "Collaborative Filtering by Mining Association Rules from User Access Sequences" Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration. (2005)

[17] Amy Shi1, Allen Long2, and David Newcomb3 "Enhancing e-Business through Web Data Mining"

[18] Usama Fayyad, et al. "From Data Mining to Knowledge Discovery in Databases" Ai Magazine (1996)

[19] Badrul Sarwar, George Karypis, Joseph Konstan and John Riedl "Item-based Collaborative Filtering Recommendation Algorithms" World Wide Web (2001)

[20] Graham J. Williams and Zhexue Huang "Modeling the KDD Process" CSIRO Division of Information Technology Data Mining Portfolio (1996)

[21] http://www.x-solutions.poet.com/eu/newsevents/glossar/#ECommerce

[22] http://www.thearling.com/text/dmtechniques/dmtechniques.htm

[23] Adomavicius, G. and Tuzhilin, A. "Using Data Mining Methods to Build Customer Profiles" IEEE Computer Society.(Feb 2001)

E-commerce Site Report

http://Beautysassy.com

By Eunjeong Cha (683998)

September 31, 2006

Thesis Report Submitted by EunJoeng Cha, Department of Computer Science

1. Site Description

This is e-commerce to sell wigs and hairpieces. This company has their own off-line business as well. When I started this project they wanted to build not only web-site but also recommendation system. One more request was they would like to reflect their own business know-how.

I made recommendation items and best selling products in page category and product category for helping user find product what they want to buy.

This Website just launched a month ago. We don't have enough data for recommendation based on customer. So I decide to use sale person personal opinion from his/her long experience. I added three fields to apply in database table and then I can use that as filter in each page.

The Goal of this site implement is intelligent recommendation system using data mining algorithm in combined data such as web log, purchase history and experience. I expect the result will be more reliable than now.

I will present this site design and implement and working process. Also I will show what I wanted to implement in each page.

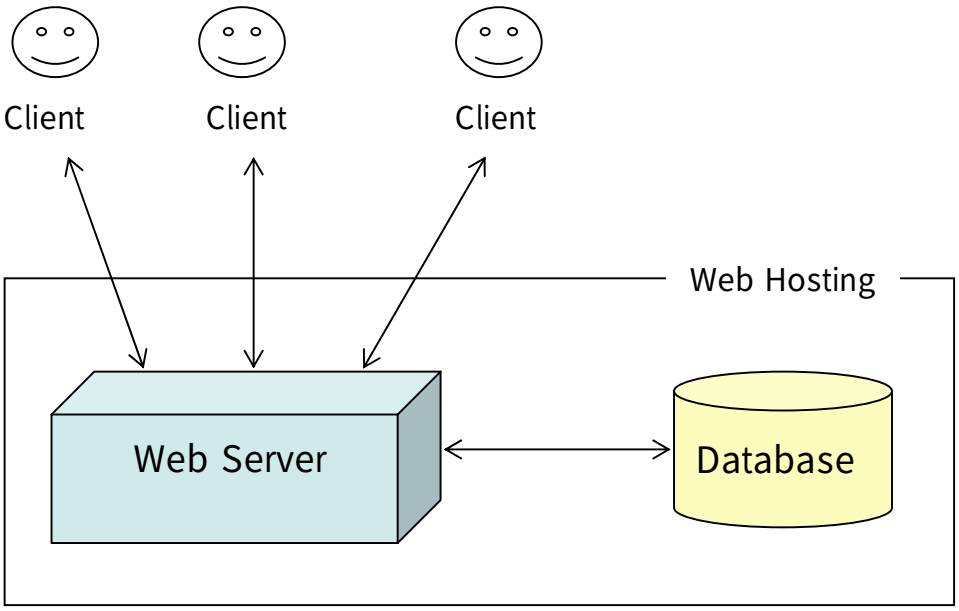
2. System Architecture

Operating System : Windows Advanced Server 2000

Database Server : MS-SQL Server 2000

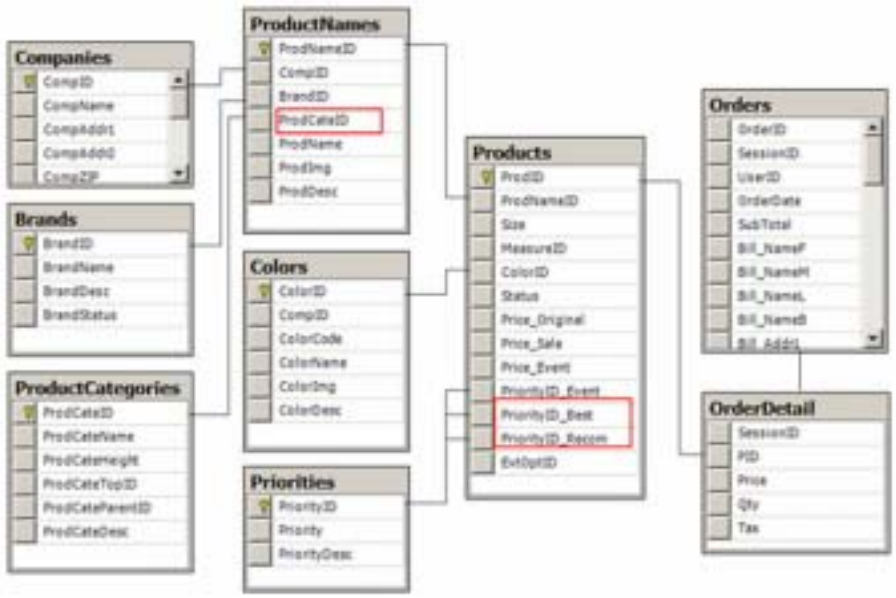
Web Server : IIS 5.0

Develop language: ASP .NET with C#, Java Script and Html



This site is common 3-tier architecture like above figure. All the programming code, front-end and back-end, is in web server and database is divided from web server.

3. Database Design



This is database design for this e-commerce site. I have two different category tables in database design. One is page category and another is product category.

There is one table for product category, ProductCategories. This table contain all the product category, category depth and category parent ID. Below figure is table data. This is used for main menu in site, in other word, this site menu generated automatically based on database data. One more importance is filter when I get recommendation items in web page. TopCategoryID and ParentID are designed to extract data in the future.

Table - dbo.ProductCategories

ProdCateID	ProdCateName	ProdCateHeight	ProdCateTopID	ProdCateParentID
1	HairPiece	1	NULL	NULL
2	Wig	1	NULL	NULL
3	Human Hair Wig	2	2	2
4	Synthetic Wig	2	2	2
5	Human Hair Weave	2	1	1
6	Human Hair Braid	2	1	1
7	Syntetic Weave	2	1	1
8	Synthetic Braid	2	1	1
9	Human Hair Half ...	2	2	2
10	Synthetic Half Wig	2	2	2
11	Human Hair Pon...	2	1	1
12	Synthetic PonyTail	2	1	1
13	Human Hair Shor...	2	1	1

Here, I will explain page category in Priorities table. This information make data range narrow. For instance, I am looking product detail page I cannot see product which is selected as home recommendation items.

Table - dbo.Priorities

PriorityID	Priority	PriorityDesc
1	home	Home Page
2	category	NULL
3	product	NULL
4	NONE	NONE
NULL	NULL	NULL

Now, you can see products table data how I saved it. Each product has 3 different priorityID's. In this, I mentioned PriorityID_Recom which has category level information page.

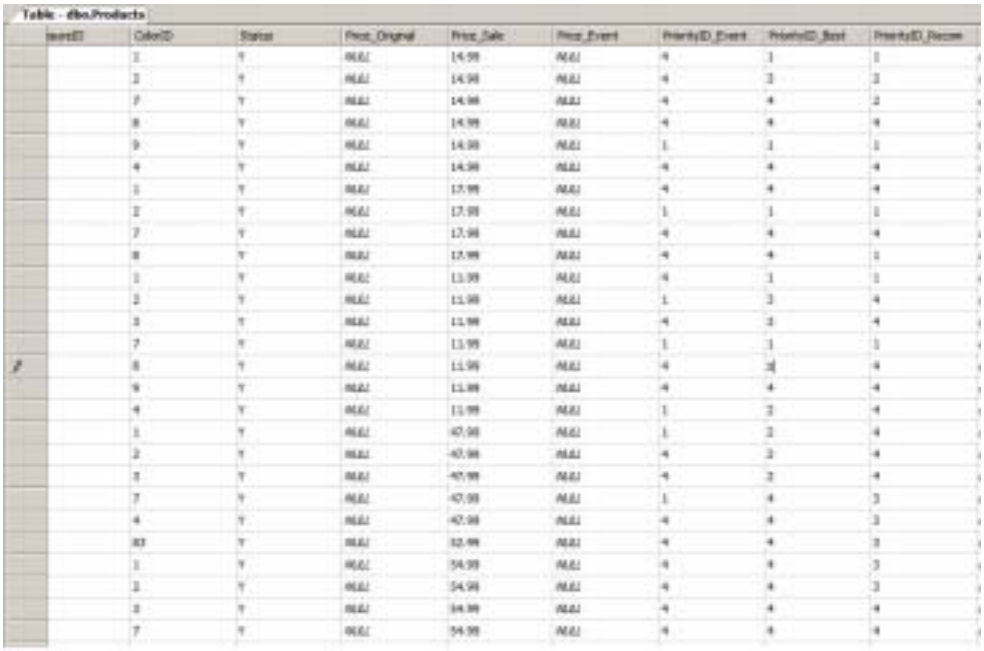
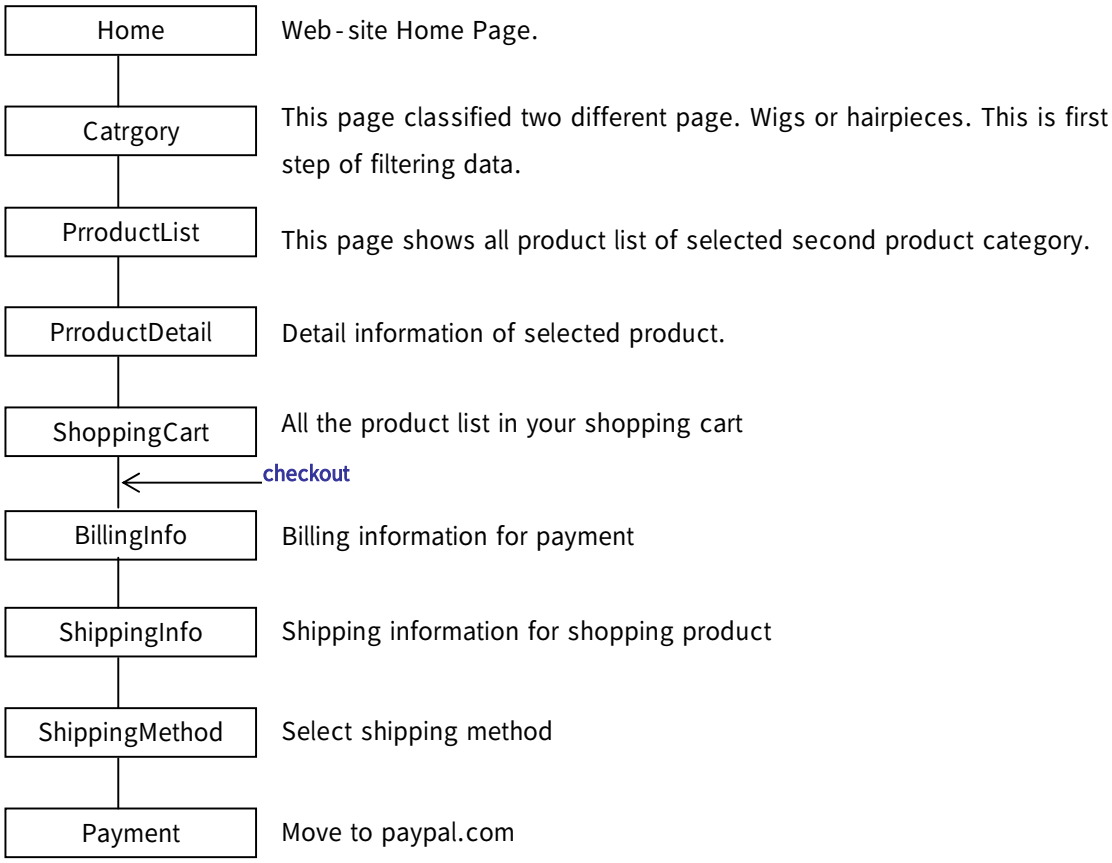


Table - dbo.Products
ColorID
Status
Price_Original
Price_Sale
Price_Event
PriorityID_Event
PriorityID_Best
PriorityID_Recom

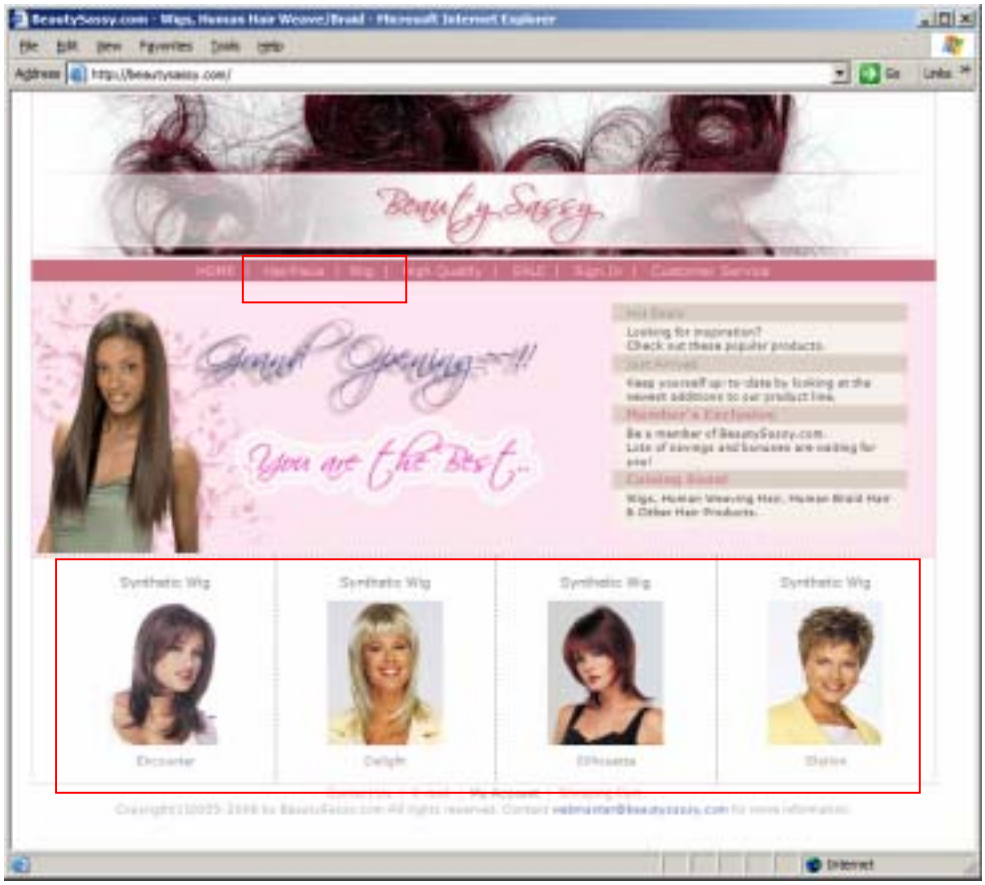
4. Web-Site Process



This figure shows site whole process. I will explain all pages what I considered when I designed this site and what I am implementing more with screen shot.

5. Screen shot

**** HOME**



First, I want to mention Main menu. This menu box generated automatically by database data what is “ParentsID is Null” in PoductCategories. “ParentsID is Null” means that category is top category.

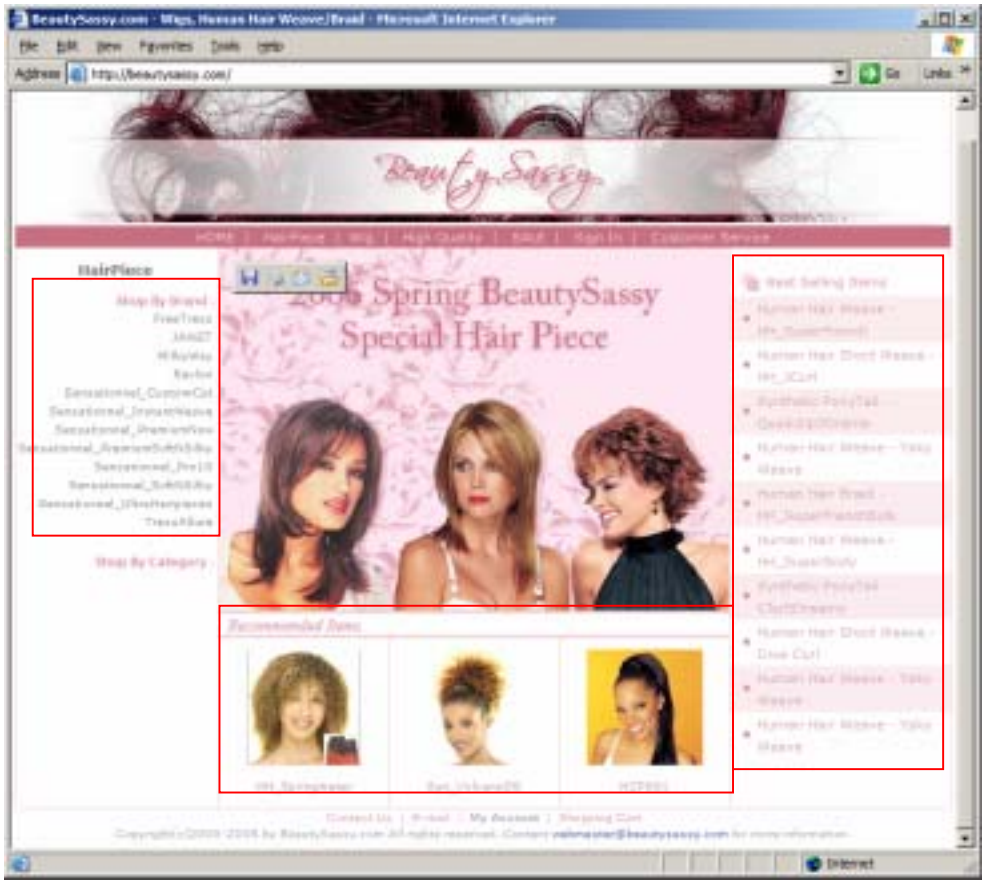
Second, This page shows 4 recommended items. Below box is pseudo query.

```
Select Top 4 [Fields..]
from [Table_Name,...]
where .....
and Products.PriorityID_Recom in (1,2,3)
order by NEWID()
```

When we load home page these 4 products are changed because of “order by

NEWID()" that is random function in MS-SQL Server. In this page, I don't need to use product category for filtering data because home shows only 4 items within all data and all categories.

**** Category**



First, I present about recommended items. Below box is pseudo query. Here is one more condition for filtering product category, also use random function. This page also show different items every time when you reload this page.

```
Select Top 3 [Fields..]
from [Table_Name,...]
where .....
and Products.PriorityID_Recom in (2,3)
and ProductCategories.ProdCateTopID = '"+CateID+"'"
order by NEWID()
```

Second, you can see best selling items at right side. That is same concept with recommended items but just used different field like “PriorityID_Best”.

Third, there is one browsing menu at left side in this page. This menu populated dynamically and this spread menu work with java script.

Fourth, I will give a line of web log and explain what kind of information I can get. If I have a lot of web history, I can discover knowledge such as which category is more interesting when people see category page. And then I can add more products or more promotion.

```
2006-08-23 08:22:50 66.249.72.103 - W3SVC188 IIS39 72.41.209.55
80 GET /yu/category.aspx CateID=1 200 0 19517 265 187 HTTP/1.1
```

** ProductList

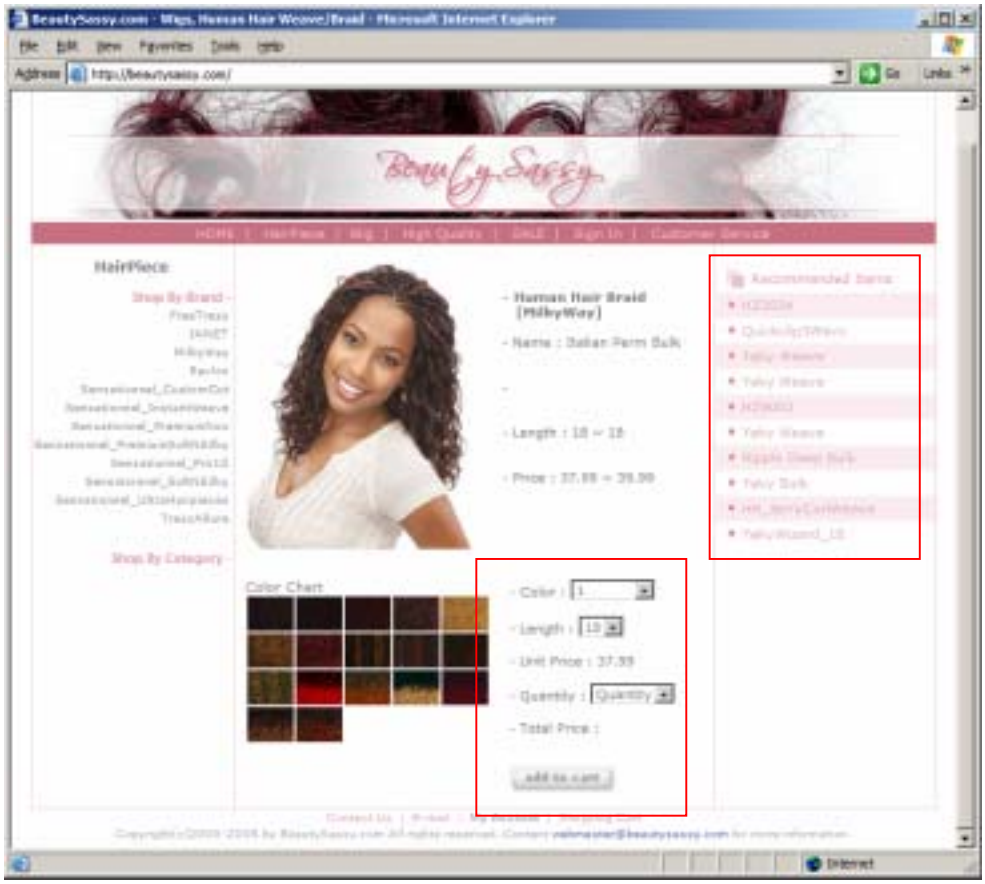
This product list is populated by category id and sub category id.

```
2006-08-23 08:22:50 66.249.72.103 - W3SVC188 IIS39 72.41.209.55
80 GET /yu/productList.aspx View=Brand&SubID=1&CateID=1 200 0
```

Here is one more web log example line. If View is Brand then subID is brandID, if View is Cate then subID is CategoryID. I can get rule from this like what kind of brand is interested when people shop by brand or what kind of category is interested when people shop by category.

That will be implemented in next step.

**** ProductDetail**



In this page, I am concerned of two things: one is recommendation items and another is how to get price.

There are recommendation items at right side of this page. It is same logic with other Recommendation items.

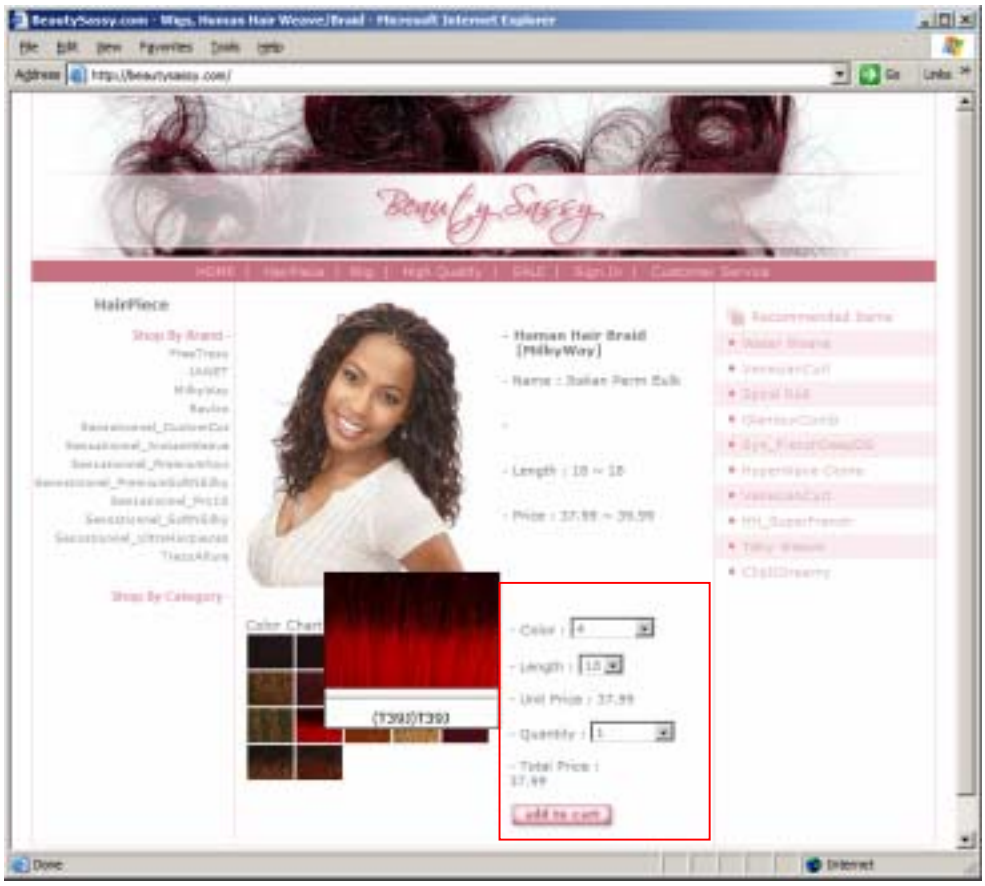
```
Select Top 10 [Fields..]
from [Table_Name,...]
where .....
and Products.PriorityID_Recom in (3)
and ProductCategories.ProdCateTopID = '"+CateID+"'"
order by NEWID()
```

Here is NameID which can decide product. This product is special feature about the price. This price of hair pieces depends on length and color. Long hair charge

more money in same hair. So First customer decide product and then can decide color and length. NameID is important information to suggest items.

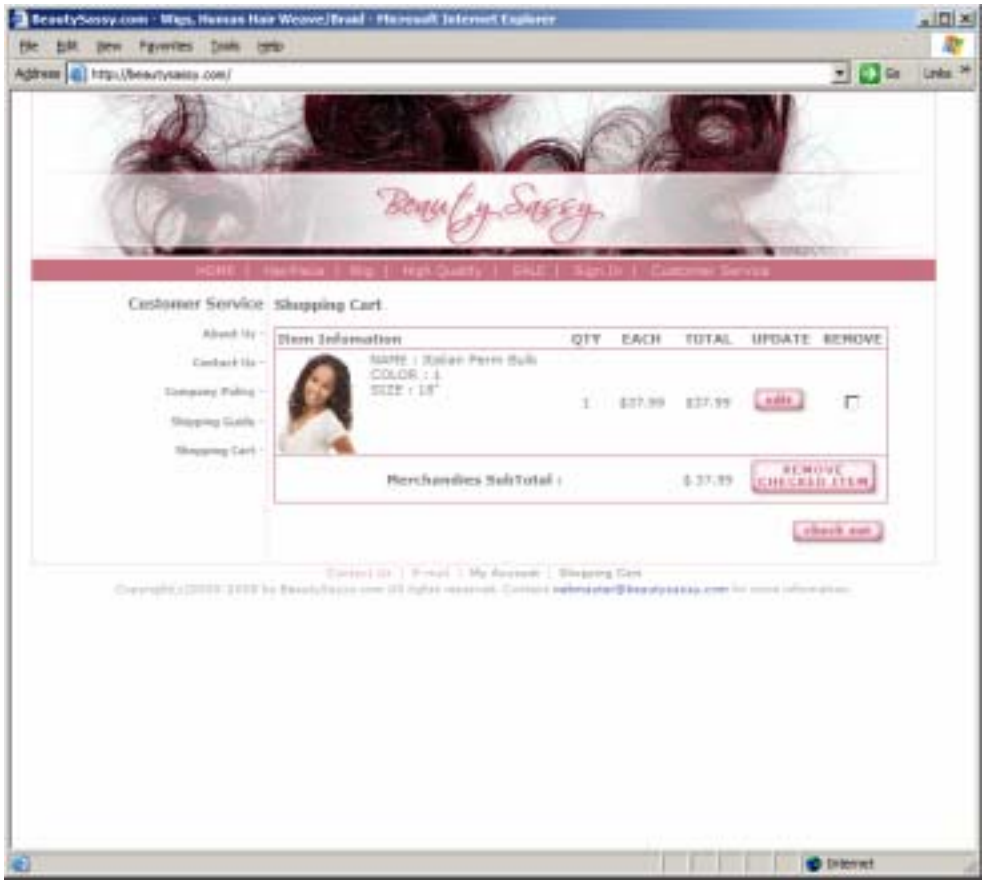
```
2006-08-23 08:22:50 66.249.72.103 - W3SVC188 IIS39 72.41.209.55

80 GET /yu/productDetail.aspx CateID=1&NameID=54 200 0 19517 265
```

Next, I mentioned about different price of product. Customer finally can order after select color and length so I made button disable first. When people input all information, then button will be enabling like below figure.



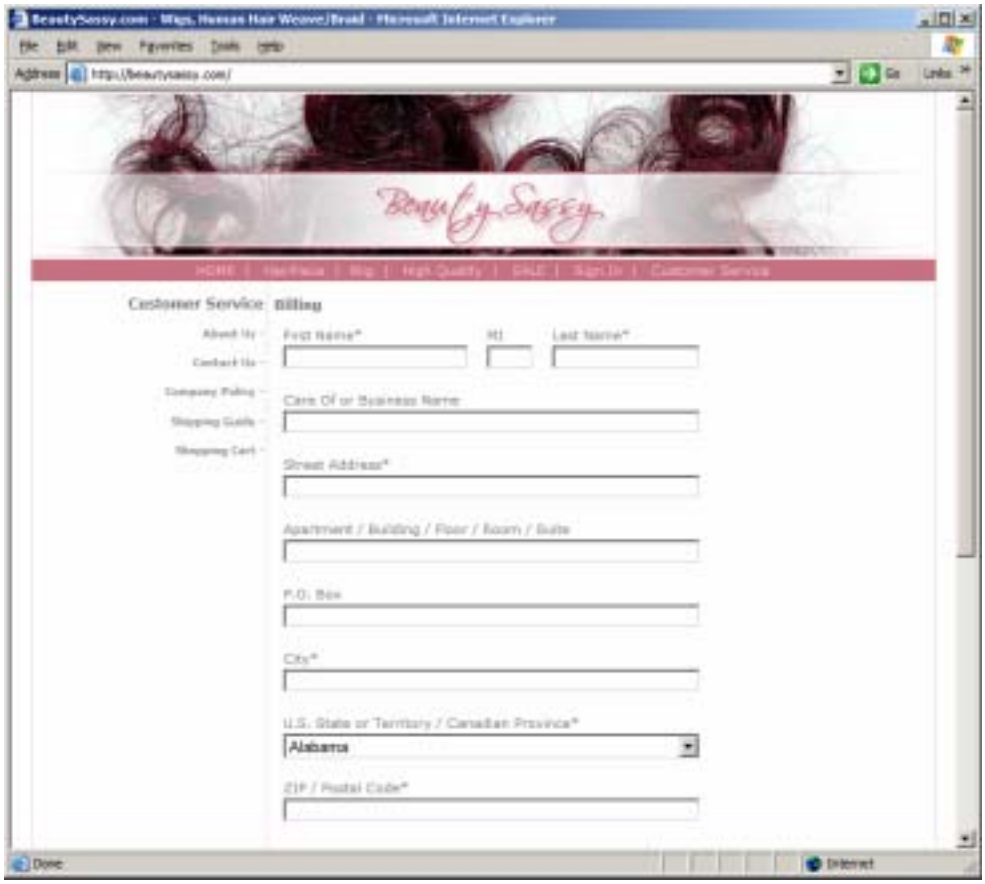
Clients can see hair color in detail as roll over mouse on color as above figure. This is implemented by javascript. This color is only for this selected product, in other words, if customer choose other product then other color list will be shown.

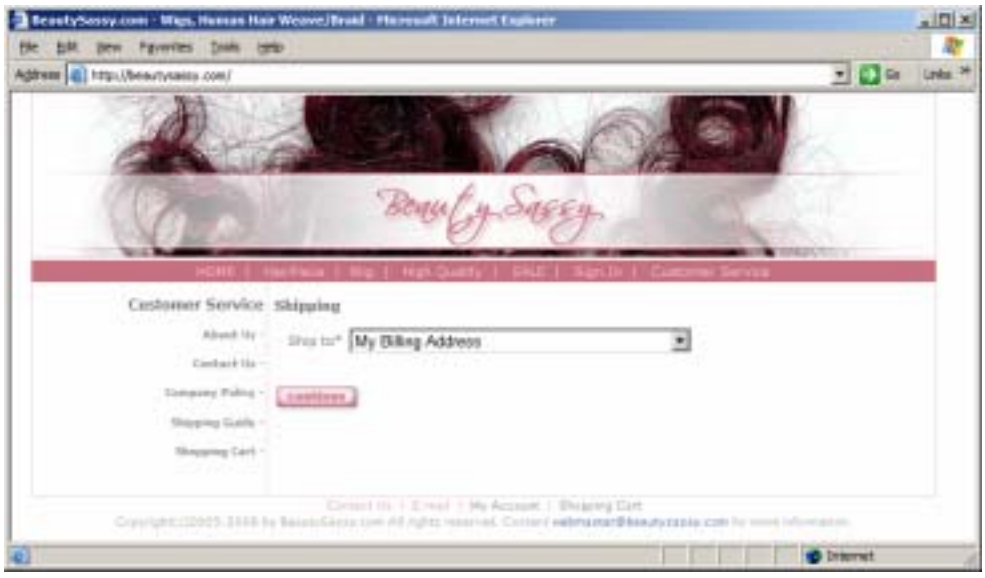
**** ShoppingCart**



I am using session for shopping cart and all order information. It will be history of purchase in database, order and orderdetail table.

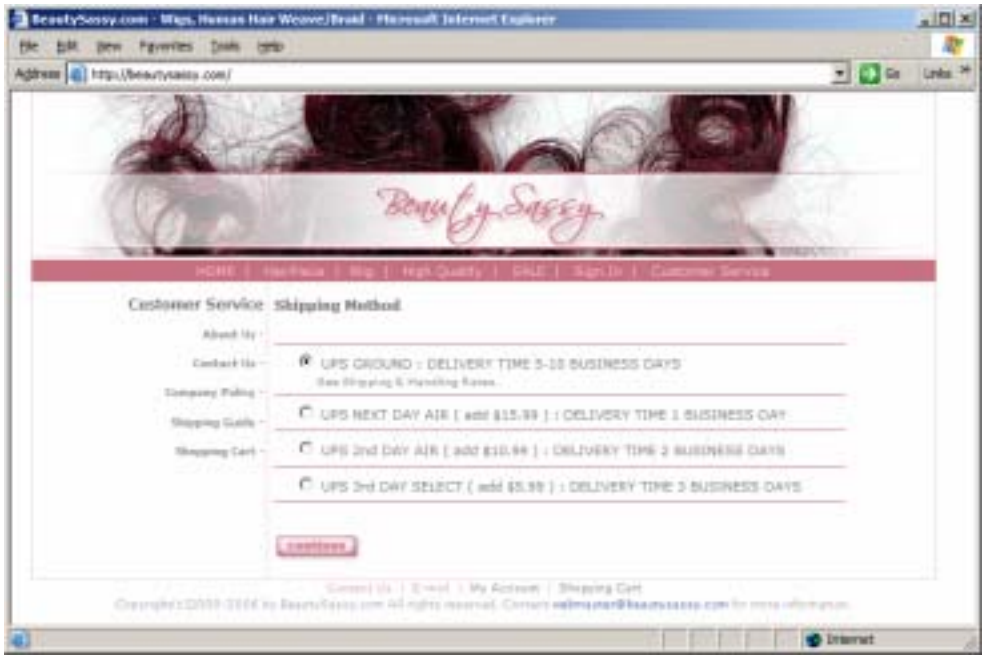
** Billing and Shipping Infomation



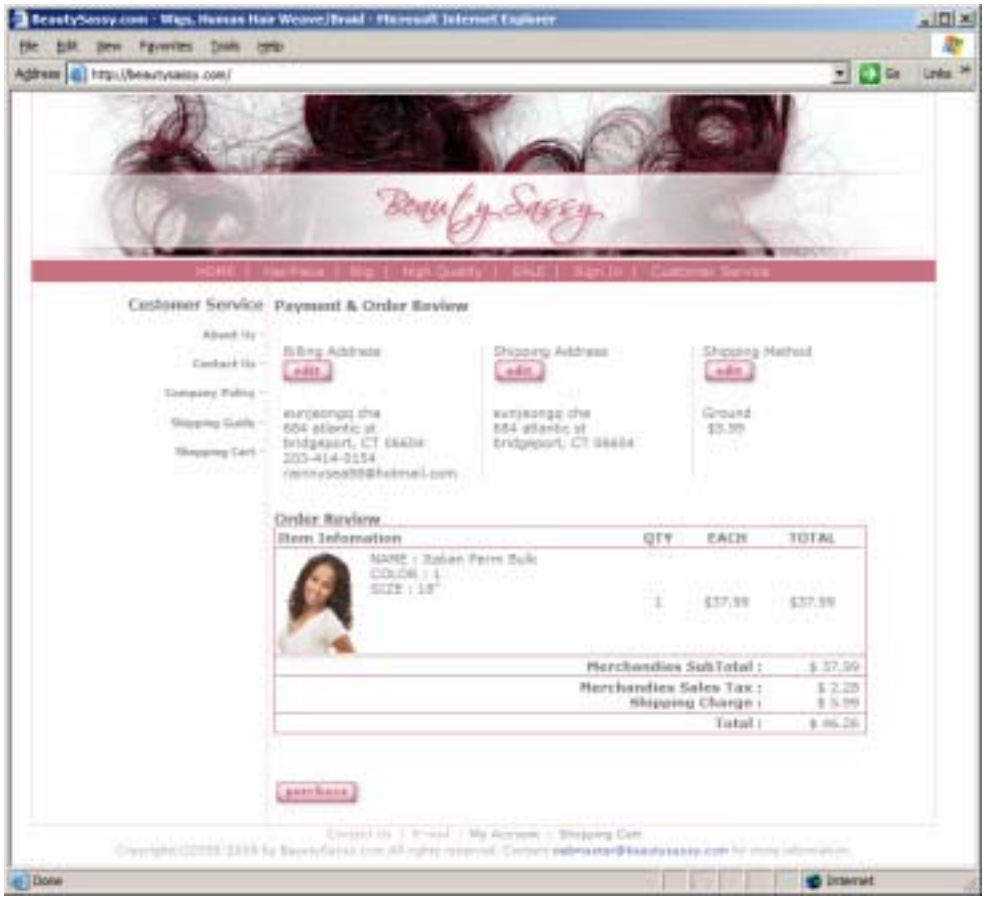


These are common shopping process with other sites. These pages are also using session.

**** Sipping Method and Summary**



This site supports 4 different shipping methods with different fare. When you continue shopping, this program will calculate shipping fare.



First, This page shows order summary of customer and then they can modify wrong or missed data.

This “purchase” is last process of this web site. When people purchase product, web site save purchase history in database, read from session variable and then send information to paypal. In this case I was worried about lost internet connection so I did one more thing after paypal process. When process is coming back from paypal I updated order information which I got from there. The field is PaypalCode in Orders table.

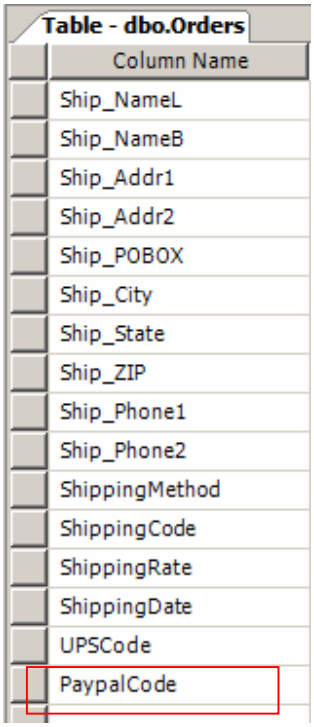
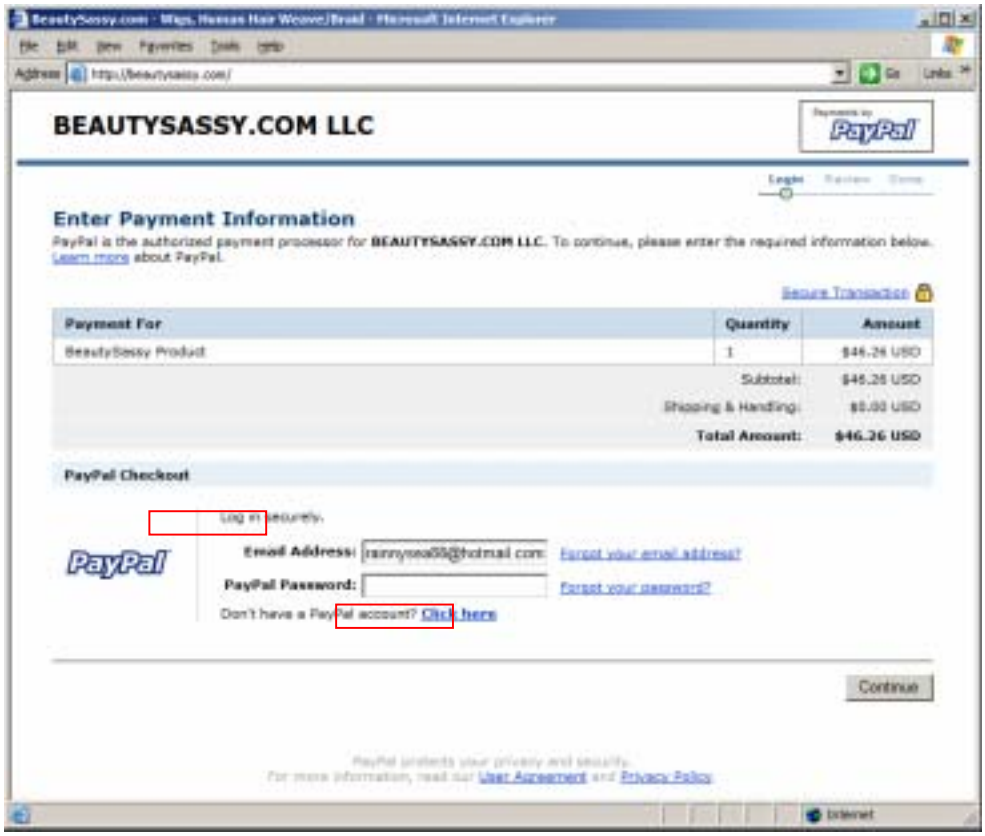


Table - dbo.Orders
Column Name
Ship_NameL
Ship_NameB
Ship_Addr1
Ship_Addr2
Ship_POBOX
Ship_City
Ship_State
Ship_ZIP
Ship_Phone1
Ship_Phone2
ShippingMethod
ShippingCode
ShippingRate
ShippingDate
UPSCode
PaypalCode

**** Payment (paypal.com)**



I didn't implement payment system in this site because of security problem. This site supports to pay member and non-member both of them. When I designed this E-commerce, we discussed about security problem and then I came to conclusion not to implement payment system in web site and use reliable company.

This web site returns back to beautysasy.com after finishing pay. I can get purchase history from two other place, one is beautysassy website and another is paypal website. So I can compare If something happen like lost data.

6. Further Implementation

I presented my all work, system architecture, database design and page process. This recommendation system in this website designed automation operation as well as manual operation as their request. This recommendation system helps users find items they want to buy and is available cross-selling in product detail page. Now I am working on extract information from web log and get knowledge discovery using data mining algorithms. This system is going to be improved mining technique with more various data such as web log, selected items and purchase history and searched optimized recommended items.