MINING PROGRESSIVE USER BEHAVIOR FOR E-COMMERCE USING VIRTUAL REALITY TECHNIQUE

A Thesis

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

---

by

MING-CHANG CHEN

Dr. Chi-Ren Shyu, Thesis Supervisor

DECEMBER, 2007

The undersigned, appointed by the dean of the Graduate School, have examined the thesis entitled

MINING PROGRESSIVE USER BEHAVIOR FOR E-COMMERCE USING VIRTUAL REALITY TECHNIQUE

presented by Ming-Chang Chen,

a candidate for the degree of Master of Science

and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Chi-Ren Shyu

Dr. Jeffrey Uhlmann

Dr. So-Yeon Yoon

To my dearest parents, who have raised and educated me with their unconditional love & care.

To my loving wife, who loves me for who I am, and supports me in my life, for being a good friend and family forever.

ACKNOWLEDGMENTS

My greatest gratitude is to my thesis advisor, Dr. Chi-Ren Shyu, who has been providing me with professional directions and solid academic training patiently. He has always been inspiring and encouraging throughout my graduate study. His intellectual guidance and attitude towards life have not only led me to achieve a higher goal in academic but also inspired me to have greater expectations of myself. Although the thesis is extremely challenging, Dr.Shyu's warmest smile and sense of humor during the discussions made it an easier task for me.

Another thanks goes to my thesis committee members, Dr. So-Yeon Yoon and Dr. Jeffrey Uhlmann for their insightful comments and encouragement. I would also like to thank Dr. Hyunjoo Oh from University of Miami for providing me the opportunity to participate in this project. Their efforts and time are sincerely appreciated. Also, some credits go to my lab mates from the Medical and Biological Digital Library Research Lab for their help and friendship.

Finally, I am grateful to everyone who has assisted me during my studies. Thank you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

MINING PROGRESSIVE USER BEHAVIOR FOR E-COMMERCE USING VIRTUAL REALITY TECHNIQUE

Ming-Chang Chen

Dr. Chi-Ren Shyu, Thesis Supervisor

ABSTRACT

In the past decade, the virtual reality (VR) technique has been becoming a popular marketing tool on e-commerce websites, where the consumers are allowed to interact with the products and have a vivid shopping experience simulated to the real world. Despite of the advantages, VR shopping is still at its early stage due to some difficulties, namely, 1) complex communication among VR components, backend relational databases, and web services, 2) meaningful interpretation of mined patterns into the user behavior through VR tools, and 3) vast amount of information to mine and efficient summarization of the results to guide user's navigation.

To overcome these obstacles, several techniques from the fields of Information Retrieval and Data Mining & Knowledge Discovery have been adopted and extended for the development of our system that is able to recommend products computationally according to the user's preference. These techniques include ranking in a Vector-Space Model to profile user's behavior based on their demographic information and mining Association Rules to analyze user's choices as well as real-time navigational behavior. This system can adapt itself to meet each individual's interest progressively. It is our ultimate goal to provide a general framework that can be applied to any web-based e-commerce applications where customers can find the most fitting products efficiently and effectively under a virtual reality environment.

Chapter 1

Introduction

1.1 Motivations

The World Wide Web, a virtually infinite storage space, contains huge amount of data and information. In such a huge and global information center, the users are provided with a wide variety of resources and messages. It also enables the marketers to easily reach the customers and promote their brands or products by offering vast product information and options. As consumers' acceptance of online shopping has grown, retail e-commerce has evolved to encompass various goods and services, and consumers continue to expand their online shopping experience into new product categories (Cole, Suman, Schramm, Lunn, & Aquino, 2003). There is no doubt that the e-commerce has become an essential marketing strategy in business. However, due to the increasing number of choices to make and vast information to process, online shopping could be an exhausting and overwhelming experience for the online customers. One solution to this information overload problem is to tailor the website to the users' needs and provide the consumers with personalized information.

1.2 A need for Adapting Recommendation System

The CEO of Amazon.com, Jeff Bezos, once said, “If I have 3 million customers on the Web, I should have 3 million stores on the Web.” His words are simple but powerfully point out the importance of web personalization. Web personalization can be defined as “any action that adapts the information or services provided by a Web site to the needs of a particular user or a set of users, taking advantage of the knowledge gained from the users’ navigational behavior and individual interests, in combination with the content and the structure of the Web site (Eirannis & Vazirgiannis, 2003).” In the real world, it is nearly impossible to build a brick-and-mortar store for every single customer. However, on the Web, by the means of personalized techniques and recommendation systems, every customer can virtually shop at his or her own store with tailored information and product arrays according to the individual needs and preferences.

Recommendation systems have been widely used as a customization tool to predict consumers’ interests in the products and have been increasingly popular on the Internet. E-commerce sites, such as Amazon.com and eBay, use recommendation systems to provide consumers with tailored information based on demographic information, past shopping behavior, ratings, or the top overall sellers in order to help consumers decide which products to purchase (Schafer, Konstan, & Riedl, 1999). E-commerce websites with recommendations systems create a fantastic shopping environment in which every customer is treated individually with specific information. Recommendation systems can be considered as every customer’s personal assistants or consultants while shopping as they can dynamically re-arrange the display of the other products for the next choice. Rather than being exposed to the fixed webpage, the users are provided with flexible

product presentation and suggested with suitable items. With the assistance of recommendation systems, not only can consumers spend less time surfing or looking for the products that interest them but also find the potential items that they might not know yet.

The major objective of this project is to propose a hybrid recommendation system for a 3D virtual reality furniture website by analyzing the users' demographic data, such as gender and income information, with a combination of weighting method imposed on the features of the chosen items to offer the recommendation lists and product presentations adapted to each customer. When shopping for expensive or high-involvement products on a website, like cars and furniture, consumers' purchase decision making is more complex and time consuming because more product messages and individual-difference moderators are involved (Putrevu & Lord, 2003). The choices of furniture depends on individual factors, such as background, tastes, and budgets, combined with a consideration of product features, such as price, comfort levels and styles. In our proposed recommendation system, several techniques, including Vector-Space Model from Information Retrieval and Association Rules from Data Mining, are adopted in order to generate a hybrid self-learning recommendation system. With this system, the customers can very likely find the most suitable furniture products in the shortest time without spending much effort.

Chapter 2

Literature Review

2.1 3D Virtual Reality in E-commerce

The advanced technology of virtual reality has been recently applied on the e-commerce website where the consumers can inspect the product from different angles and manipulate it to create visual images. It also enables the customers to mix and match products from different categories to see if they fit together. Such virtual experience, defined as “psychological and emotional states that consumers undergo while interacting with products in a 3D environment” (Li, Daugherty, & Biocca, 2001), has been found to enhance the consumers’ product knowledge, engage the customers to the website, build positive attitudes toward the brand and increase the likelihood to purchase (Li, Daugherty, & Biocca, 2002). It is suggested that the interactivity on an e-commerce website leads to more information processing (Sicilia, Ruiz, & Munuera, 2005) and can entice the customers to visit that retailer’s brink-and-motor store (Fiore & Fin, 2003). The inherent advantages of 3D virtual environments where the consumers can control over the product information and enjoy the novel experience make the e-commerce site a more powerful tool than the traditional stores and a more attractive one than the 2D sites (Edwards &

Gangadharbatla 2001; Li et al., 2002).

However, even though the 3D virtual stores outperform the traditional ones in many ways that customers can navigate at their own pace, search for the product information and interact with plenty of products, a poorly designed virtual e-commerce website can have some adverse effects on the consumer's shopping experience (Chen, Gillenson and Sherrell, 2004). Facing the huge quantity of products and details on a virtual reality store, the customers expect high quality guidance and assistance in finding their ideal products. It urges the need to personalize the e-commerce website by offering dynamic recommendations according to the individual customer's behaviors and personal needs. In this way, a virtual reality e-commerce website can enrich the users' shopping experience and achieve its maximum effectiveness.

2.2 Web Personalization & Recommendation Systems

Web personalization, consisted of activities such as providing customized information, changing the webpage layouts and adapting the contents tailored to the user's need, has become an essential part of a website to enhance its compatibility and attractiveness. The recommendation system (e.g., Schafer et al., 1999) or interactive decision aids (e.g., Haubl & Trifts, 2000) can be considered as one form of personalization to facilitate in helping the users making purchase decisions.

One of the main differences between the traditional brick-and-mortar stores and e-commerce websites is the virtually infinite product presentation room or shelf-space on the Web. Unlike the traditional stores which have limited storage, the E-commerce websites provide the consumers a wide variety of options, alternatives and product

information. The diversity of product choices and the abundance of messages on an e-commerce site have led to the demand of web personalization and real-time adaptation catering to the user's need. It is because the shopping experience can be overwhelming especially when there is no assistance available in deciding what products to purchase. In addition, the effort and time spent on searching aimlessly may lead to poor quality of decision and dissatisfaction of the consumers (Chen et al., 2004). Therefore, to find the ideal products in mind effectively and efficiently, online customers not only look for the suggestions from their peers, and editorial picks (Smith, Menon, & Sivakumar, 2005) but also heavily count on the real-time recommendation systems featured on the e-commerce websites (Haubl & Trift, 2000; Senecal & Nantel, 2004), such as Yahoo!, eBay and Amazon.com (Ansari, Essegaier, & Kohli, 2000).

The recommendation system, a technology that enables the e-commerce websites to predict a user's interest for a product and provide personal shopping assistance according to individual background information, preferences and needs, have become a reliable and influential tool for the customers (Haubl & Trifts, 2000; Senecal & Nantel, 2004). Evidently, more and more e-commerce sites resort to recommendation systems to interact with the customers and hopefully increase sales. There are a number of techniques involved in web personalization and recommendation systems, such as user profiling, Web usage mining log analysis, and information retrieval (Eirinaki & Vazirgiannis, 2003). The following paragraphs provide overview of some of these techniques.

User Profiling

The idea of user profiling is to divide the total users of a database into several subsets based on their similar characteristics or navigating behaviors. For each subset, the prediction is made from the averaged opinions across the whole users (Eirinaki & Vazirgiannis, 2003). In other words, when the recommendation system makes a suggestion for a user, the decision is made upon the majority in that subset profile. Sometimes, one customer can be represented partially in several subsets; then, the prediction will be weighted by degree of participation (Schafer, Konstan, & Reidl, 2001). Once the profiling process is complete, the performance of analyzing can be enhanced because the size of users in each group is smaller. Most of the time, user profiling is considered as a first step to reduce and filter a large size of users.

Association Rules

Association Rules (Agrawal, Imielinski, & Swami, 1993; Agrawal & Srikant, 1994) in data mining is a technique that analyzes the correlations and patterns between sets of items. It has been widely used for the past decade, mostly by retails stores, to find out the preferences across the products and consumers' purchase behaviors. Association rules make the predictions based on the products they already chosen. The number of possible association rules grows along with the increasing number of products. With two user-defined parameters, support and confidence, the number of association rules to show can be restrained. This method is commonly used by large population rather than individuals and is more suitable for static information than rapidly changing data format.

Web Usage Mining

Web usage mining is the process of applying data mining techniques to discover the consumers' behavioral patterns based on web data. The overall process of web usage mining can be generally divided into two main tasks: data preparation and pattern discovery (Eirinaki & Vazirgiannis, 2003). As a user's online activity always leaves a virtual trail, such as how long he/she has been on a page or a site, what mouse movement has occurred, as well as the content of every incoming and outgoing messages. Therefore, by analyzing and finding out the relationships between/among these web data, the recommendation system can determine the user's preferences and further predict or recommend the products that might interest him/her (Cho, Kim, and Kim, 2002).

2.3 Types of Recommendation Systems

A good recommendation system can increase sales of an e-commerce website because it turns browsers into buyers, builds rapports, enhance brand loyalty, and increases cross-sell (Schafer et al., 2001). Ansari and his colleagues suggest that recommendations can be generated at least based on the following five sources: "(1) a person's expressed preferences or choices among alternative products, (2) preferences for product attributes, (3) other people's preferences or choices, (4) expert judgments, and (5) individual characteristics that may predict preferences" (Ansari et al., 2000). Such information can be acquired from a user's explicit profiles, real-time navigational behaviors, ratings, as well as the product choices, and can be stored organizationally in the database. By knowing the users well, providing immediate interactions and adapting dynamically to different needs, recommendation systems can enhance the customers' purchase

experiences and ultimately benefit the e-commerce website (Schafer et al., 1999). In the literature, a large number of recommendation techniques have been developed and applied (e.g., Ardissono, Goy, Meo, Petrone, Console, Lesmo, Simone & Torasso, 1999; Cho et al., 2002; Krulwich, 1997; Lee, Liu, & Liu, 2002; Lekakos & Giaglis, 2007). These systems can be generally categorized into four: demographic filtering approach, collaborative filtering approach, content-based filtering approach, and hybrid approach.

***Demographic filtering*:**

Similar to the traditional marketing segmentation methods, demographic filtering approaches categorize the customers and make recommendations based on their explicit personal information and demographic data. These data are "objective facts" such as record data, geographic data, personal attributes, psychographic data and lifestyle (Kobsa, Koenemann, & Pohl, 2001). Normally, the users' demographic or background information is obtained during the registration process and analyzed with techniques such as classification, clustering, user profiling, information retrieval, or data mining (Ansari, Essegaier, & Kohli, 2000; Cole, Suman, Schramm, Lunn, & Aquino, 2003). This approach attempts to discover the hidden correlations between the stereotypical descriptions of a user (such as gender, age, occupation, ethnicity, and income) and the features of the selected item in order to create an initial screening model for the user (e.g. Ardissono & Goy, 2000; Ardissono, Goy, Meo, Petrone, Console, Lesmo, Simone, & Torasso, 1999; Kruwlich, 1997; Pazzani, 1999).

Although this approach has been widely employed in the marketing to segment the customers and predict their interests, there are some critical shortcomings.

For example, because it classifies users and makes recommendations based on statistic descriptions and stereotypical judgment, the filtering results may be too generalized and neglect individuals' specific needs, like looking for a gift. Also, the user's registered information is not updated and could become meaningless over time. Therefore, demographic filtering method is rarely used as an independent tool in recommender systems (Burke, 2002; Montaner, Lopez, & Rosa, 2003; Wei, Moreau, & Jenning, 2005). However, when integrated with the user's real-time behaviors by combining other filtering techniques, it can function as an effective fundamental screening tool to initially rule out redundant information (e.g., Vozalis & Margaritis, 2004; Ardisso et al., 1999; Ardissono & Goy, 2000).

***Content-based filtering*:**

Content-based filtering approaches make recommendations simply based the features of the user's previously evaluated items. This method is based on the assumption that "a user's previous preferences or interests are reliable indicators for his/her future behavior (Lekakos & Giaglis, 2007)." In other words, this kind of systems tends to recommend items similar to the one that the customer has purchased or shown interests. However, predictions based on this method are too subjective and the variety of choices provided to the users is restricted (Balabanovic & Shoham, 1997; Wei et al., 2005). If the recommendation system can only suggest the items with the same or relevant features that already highlighted in a user's profile, then he/she may not have the chance to be introduced the items that he/she has not noticed yet. Another weakness of the content-based filtering method is that it can only support certain kinds of text-based documents or

well-structured data, but not those complicated mixed information which machine cannot parse, such as movies or music (Balabanovic & Shoham, 1997). For text format content, it even has to ignore the text surround the images, not to mention the aesthetic related multimedia information. Therefore, this system is often used when recommending items with text or informative features, such as books (e.g., Mooney & Roy, 2000), news articles (e.g., Lang, 1995), and web pages (e.g., Pazzani, 1997).

***Collaborative filtering*:**

Collaborative filtering approaches, instead of recommending based on his/her own past choices, suggest the user potential items that are highly rated by other customers (Balabanovic & Shoham, 1997). The prediction is based on the assumption that people who have common needs or interests tend to like the same products. In other words, unlike content-based filtering, which analyzes the correlations between the current customer and the previous behaviors, collaborative filtering methods find out the relationships among the users who have shown similar interests.

Although the collaborative filtering can solve the problems of contend-based filtering, inevitably, a pure collaborative filtering certainly introduces its own weaknesses. First of all, since the system recommends items based on other people's interests and ratings, it is not possible to recommend a new item until this new item has been introduced to some users in some other ways. Another problem is that it requires a lot of people to participate. If there is a special user who has a totally different taste compared to the majority of the population, then the suggestion based on other people's rating will be a poor recommendation to him/her

***Hybrid approach*:**

All of the three filtering methods discussed have some shortcomings and rarely can work independently to make high quality recommendations. Therefore, several hybrid approaches have been proposed in order to integrate different approaches to complement each other and make the most suitable recommendations (e.g., Burke, 2002; Lekakos & Giaglis, 2007; Pazzani, 1999; Balabanovic & Shoham 1997). Pazzani (1999) combines collaborative, content-based and demographic filtering approaches to build a system to recommend restaurants. Balabanovic and Shoham (1997) announced a hybrid system combining content-based and collaborative recommendation.

2.4 The Proposed System

In the literature, a number of recommendation systems have been proposed but rarely are designed specifically for the virtual e-commerce environment. As the fact that the e-commerce merchants are turning to 3D product presentations to stand out in such a competitive environment, the scarcity of literature in recommendation systems targeting virtual e-commerce website has evoked the need to create an adapting system focused on the 3D virtual reality shopping environment

The current thesis presents a progressive adapting hybrid ranking system for a 3D virtual reality e-commerce furniture website that assist the customers choose the products by providing personalized product presentation and recommendations in two steps. First, we utilize the user profiling approach to cluster the users into several profiles based on the concept of demographic filtering (see Chapter 4), then we adopt the vector-

space model method to determine similarity between each profile and every product (see Section 5.1). The higher level of similarity between the vectors of a profile and a product infers the higher possibility a customer would be interested in that product. Thus, by comparing the results, the system is able to make suggestions on products to any customer according to their profiles.

Although this system can customize its content to different users based on their demographic information, the recommendation is insufficient in quality. Therefore, we resort to another well-known approach association rules (see Section 5.2) to enhance the function. We adopt association rules to discover the hidden pattern of a customer's purchase behaviors based on the items that he/she has chosen. Consequently, the suggestion made by the system is based on not only the customer's profile but also on his/her personal interest. The entire procedures of this hybrid ranking system are discussed from Chapter 3 to Chapter 5, and a detailed example is demonstrated in Chapter 6.

Chapter 3

A Progressive Personalized Hybrid Ranking System

3.1 Preliminaries

The self-learning device on behalf of each user to personalize the contents, features, and structures of any applications is considered as an adapting system. Specifically, this concept was introduced to the recommendation systems in e-commerce websites to establish a vivid shopping environment for the customers. Rather than showing the customers 2D product presentation, 3D virtual stores provide more facets of the products that encourage the customers to interact and enhance their shopping experience. A previous study indicates that through the interactivities between the customers and 3D marketing signs, predictions about the consumers' final decisions can be made (Pennington, 2001). AWE3D (Adaptive Web 3D), a project built upon this idea, was proposed with a general approach for adaptive 3D Web sites (Chittaro & Ranon, 2002). Similarly, the objective of our system is to develop a 3D virtual reality e-commerce interface where the consumers can interact with the products at ease while the website

caters to the need of each individual customer.

The decision-making process in purchasing furniture is complex, involving the consideration of trade-offs between important attributes such as price, style, construction quality, comfort and other features. It also requires the consumers to spend a huge amount of time and efforts in comparing, selecting, and testing until they find the most suitable one. Therefore, for a furniture e-commerce website, where the customers are not able to touch or feel the real product, a 3D virtual reality design could amend the deficiency because it allows the customers to inspect the products with various facets and try various combinations of furniture sets. In addition, to improve the customers' shopping experience and increase their satisfaction, recommendation systems play an essential role in assisting them making high-quality decisions.

Every time a registered customer visits the website, the proposed system is able to recognize the user, learn the user's interests or preferences and adapt itself to meet the user's needs. Ultimately, this system can not only predict the items a regular customer might purchase next time but also suggest items to any first-time users based on their demographic background and real-time navigation behaviors.

3.2 Overview

To develop this shopping interface, some pretests have been conducted to gather basic information about the products and customers. To come up with a variety of product choice alternatives, we conducted two pretests using a temporary website as the stimuli development and found out the most popular features that the customers consider important by survey. Then, we compared the results to a real furniture retailer's samples

(see Section 3.3).

As to the consumers, we conducted another pretest survey to gather the participants' background information and their real-time navigational behavior to generate user profiles with demographic filtering concept (see Section 3.4). User profiling is an important step in our recommendation system. Further discussion and details are presented independently in the next chapter (see Chapter 4).

After the user profiles are set, the website start recording the customers' real-time shopping behavioral information, including the total amount of time spending on each product, all-time mouse events, such as the total number of clicks and the targets of those clicks, as well as their final decision. Inevitably, to keep track of above movements and analyze their relationships, a database to store all the information has to be created. The database design is discussed in section 3.5.

A virtual reality living room where the consumers can interact with the furniture pieces and try different displays is created with the applications from a third-party company, EON Reality Inc. (www.eonreality.com). The overall page layout of the website is shown in section 3.6.

3.3 Varieties of the Product Category

To build a reasonable and meaningful recommendation system, several considerations and steps were taken when choosing the products to be featured in the website. All of the furniture items in this virtual e-commerce website were created based on the results from the pretests so that the furniture choice alternatives can be relevant and realistic for the consumer.

In the beginning, two pretests were conducted to extract important attributes of the products. First, 88 students were asked to write down five most important features of a sofa, a chair and a table individually assuming that they were shopping for new furniture. Then, they were asked to estimate the range of price that they would be willing to pay for each item as well as the furniture set (a sofa, chair, and table). The other pretest, a total of 242 adult participants were asked to indicate the importance of the selected attributes on a 7-point scale. 1 means "not important compared to the other consideration"; 7 stands for "very critical". These attributes contain: attractive design, comfort, construction quality, quality of material, brand, fashion, range of price, durability, variety of fabrics, versatile, attractive color, pattern of fabric, and ease of maintenance.

Based on the results of these two pretests, the most important attributes were chosen (see Table 3-1). The level of prices was chosen based on the results from the pretests and industry practices as well. The validity of price ranges was confirmed by a convenient sample of retailers. The most common furniture styles that the consumers recognize are traditional, modern and casual. Each furniture model was created with three different colors. The selected attributes/levels were combined to create choice alternatives using the fractional factorial design available through SAS's JMP software (www.jmp.com). The restrictions were applied to the profile construction process in that the lowest prices ($499 and $299 for sofas; $399 and $199 for chairs; and $149 for tables) were not allowed to combine with the highest level of construction quality and the expensive materials like leather. In the end, 51 choice alternatives were created for the sofa category, 38 for chair category, and 28 for table category.

To improve this system, we have extended the original elected attributes by adding more color choices in each category of products (sofa, chair and table). In addition, for pursuing the accuracy of analyzing, another attribute, design, including features such as metal and medium wood were added in the product category (see section 5.1.2).

Table 3-1: Important features extract from two pretests for furniture choice alternatives

Attribute		Attribute Range					
Sofa	Price	$1,799	$1,299	$999	$699	$499	$299
	Comfort	▲	▲▲	▲▲▲			
	Construction quality	▲	▲▲	▲▲▲			
	Style	Traditional	Modern	Casual			
	Color/Material	Color1	Color2	Color3	...		
Chair	Price	$999	$799	$599	$399	$199	
	Comfort	▲	▲▲	▲▲▲			
	Construction quality	▲	▲▲	▲▲▲			
	Style	Traditional	Modern	Casual			
	Color/Material	Color1	Color2	Color3	...		
Table	Price	$549	$349	$149			
	Construction quality	▲	▲▲	▲▲▲			
	Ease of maintenance	▲	▲▲	▲▲▲			
	Style	Traditional	Modern	Casual			
	Color/Material	Color1	Color2	Color3	...		

▲ Fair ▲▲ Moderate ▲▲▲ Superior

3.4 Database Design

3.4.1 User Profiling Database

Every participant's demographic/ background information is the fundamental data to develop our profiling system. To obtain this information, every user is required to register a unique user ID (*userid*) and fill out a short registration form by answering some

questions. Then they can begin visiting the e-commerce website while their interactive responses are tracked and recorded. Furthermore, in order to compare their current shopping/navigation behavior with the previous ones, the number of times a particular consumer has visited, which is denoted as *sn* is recorded. A parameter called *session*, combined with *userid* and *sn* are used to distinguish the user's shopping behavior during each visiting. Table 3-2 shows some of the customers' final decisions in our database. For example, "jsimonfl-1" means it is the first time the user "jsimonfl" (*userid*) visiting this website. However, since this project is a preliminary experiment, none of the users has visited the website twice or more. Otherwise, data such as "jsimonfl-2" that indicates the user's second visit may exist.

Table 3-2: Some examples from final decisions table

Session	Sofa	Chair	Table
jsimonfl-1	sofa10	chair21	table19
rayne17-1	sofa37	chair13	table13
gatorjan-1	sofa18	chair23	table23

The user's real-time behaviors, such as how much time he/she spends on every item and how they interact with the features on left or/and right panels are also determinant data in this systems. Recording the user's interactions with the panels is done by tracking their mouse movement: *OnClick*, same as left click. In the "Action" field, from *start* and *finish*, the total amount of time each user spends on every page can be calculated (see Table 3-3). In addition, *setsofa* indicates that the users takes a look at how a furniture would fit in the VR living room; *oncheck* means the user wants to select/deselect the furniture as one of his/her favorite item; *enlarge* implies the user

Table 3-3: The navigational behavior during the shopping on the sofa page from a user

Action	Target	Time
start		2007-04-16 10:06:44
setsofa	sofa4	2007-04-16 10:07:13
setsofa	sofa39	2007-04-16 10:07:43
setsofa	sofa14	2007-04-16 10:07:53
oncheck	sofa14	2007-04-16 10:08:13
setsofa	sofa41	2007-04-16 10:08:18
setsofa	sofa45	2007-04-16 10:08:20
oncheck	sofa4	2007-04-16 10:08:24
enlarge	sofa11	2007-04-16 10:08:25
setsofa	sofa11	2007-04-16 10:08:27
setsofa	sofa27	2007-04-16 10:08:37
oncheck	sofa27	2007-04-16 10:08:42
setsofa	sofa18	2007-04-16 10:09:01
setsofa	sofa27	2007-04-16 10:09:10
setsofa	sofa27	2007-04-16 10:09:12
finish		2007-04-16 10:09:24

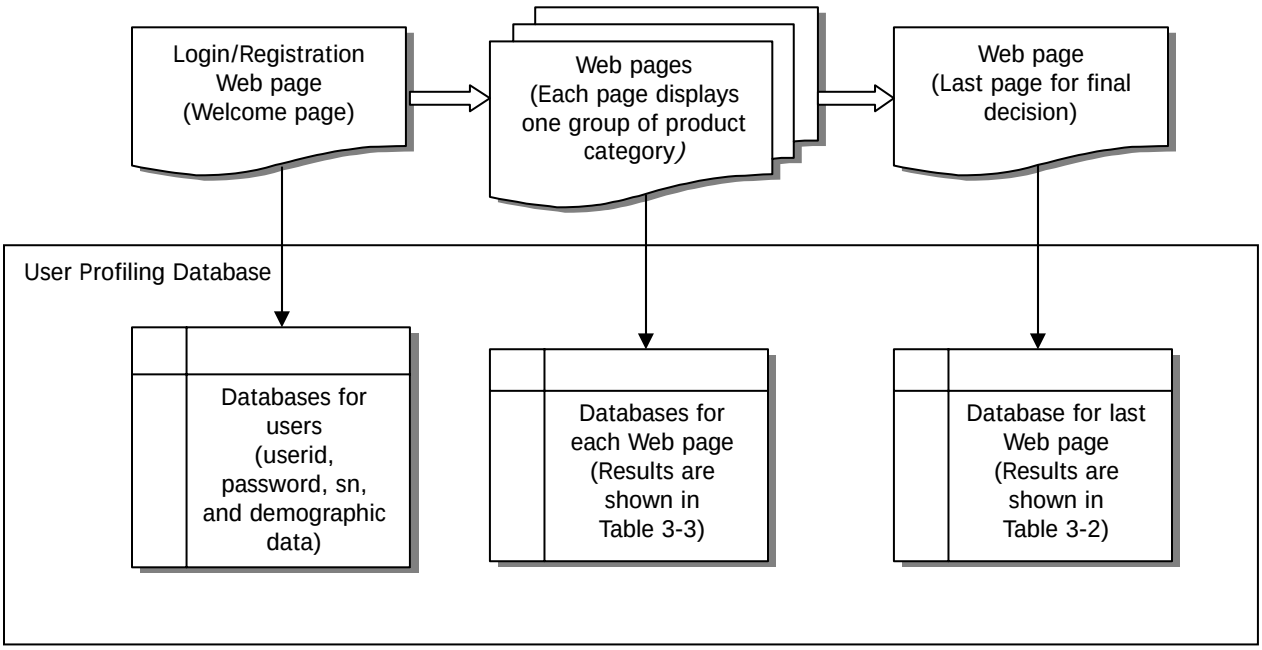

Figure 3-1: Scheme of User Profiling Database

would like to view the larger image of this furniture.

User profiling method is based on consumers' demographic information, real-time navigational behaviors, and final decisions. The scheme of User Profiling Database is illustrated in Figure 3-1, and further details will be discussed in chapter 4.

3.4.2 Vectors Database

Our hybrid ranking approach involves a lot of complex vectors and matrices calculation. The process determines how much time the system needs to generate a recommendation. If we do not compute a large part of the statistic data and exploit the database to store them in advance, a user might experience lag or feel tired of waiting for the interactions of our e-commerce website. Therefore, by calculating the data in the off-line phase, we

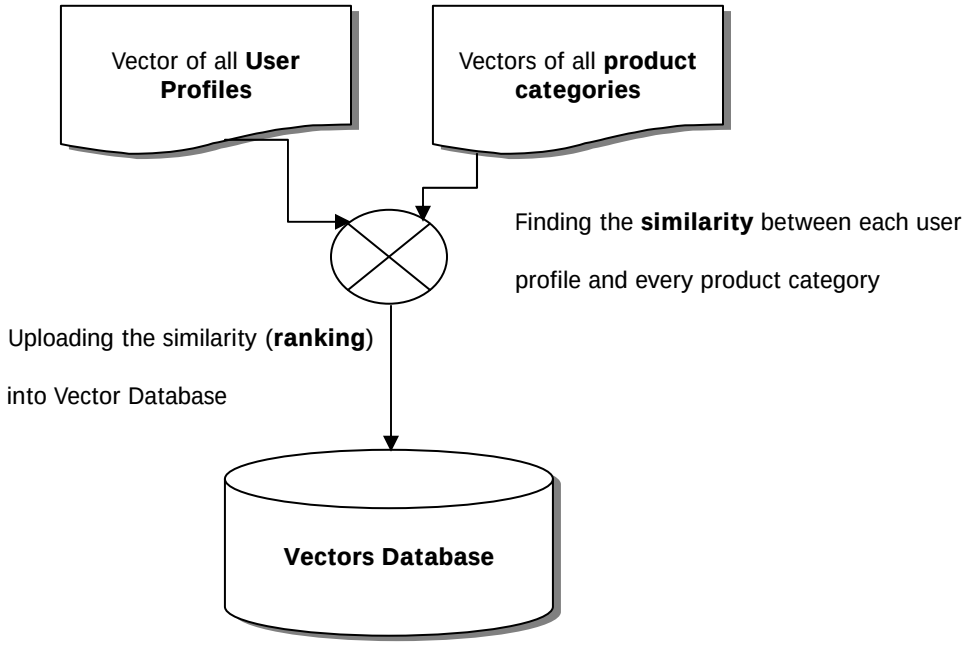

Figure 3-2: Scheme of Vectors Database

are able to accelerate our ranking system by reducing the time spent on processing.

One advantage of our hybrid ranking method is that we can compute four time-consuming jobs during the off-line stage, including 1) computing the vectors of all user profiles, 2) computing the vectors of all groups of product collection, 3) finding the similarity of each user profile and every group of product collection, and 4) uploading the above useful information (similarity = ranking) into our Vectors Database. The scheme of Vectors Database is illustrated in Figure 3-2.

3.4.3 Association Rules Database

One of the principal techniques in our recommendation system is self-learning adapting

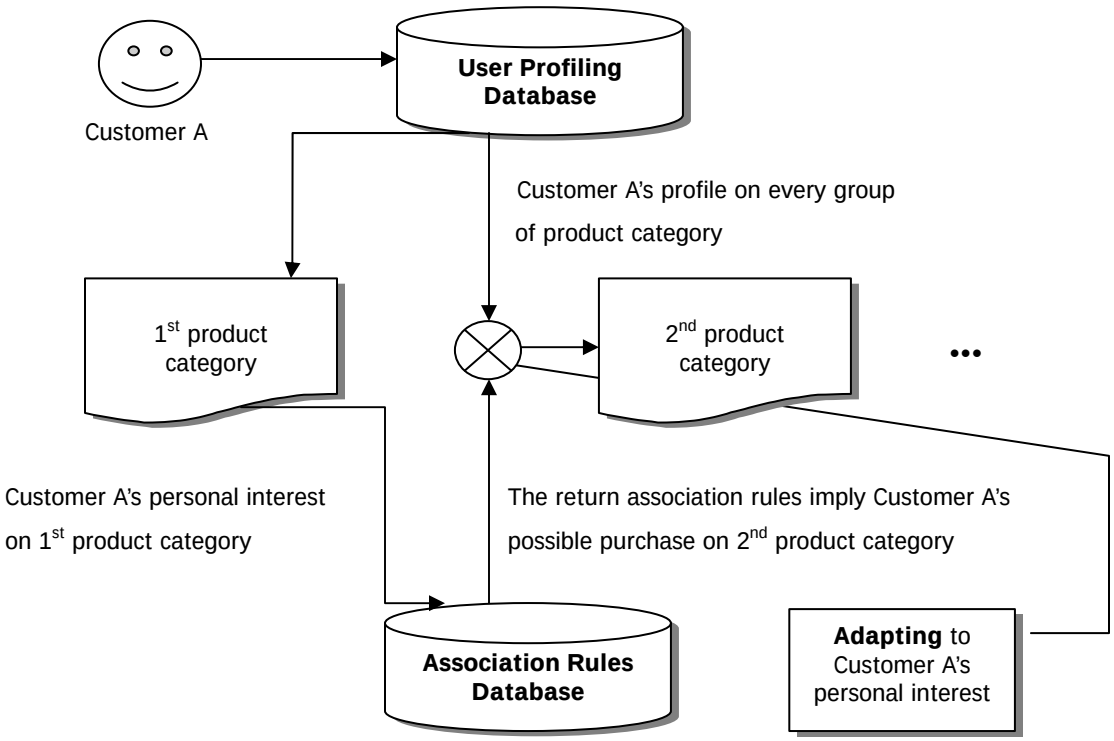

Figure 3-3: Scheme of Association Rules Database

approach. The concept of adapting system is to adjust itself to the interest of each individual. However, a recommendation, made without considering a user's personal interests and simply relying on the opinions of the majority in his/her profile, is too general and arbitrary. To overcome this issue, we adopt association rules from data mining, a method to detect correlations of specific values between each categorical variable in any set of items. By integrating this concept into our system, we discover the correlations between each choice of a particular user on every group of product collection (sofa, chair, and table) and then predict his/her preferences. The scheme of User Profiling Database is illustrated in Figure 3-3.

3.5 Website Design

Most of the websites, especially the e-commerce sites, use 2D content interface; however, 3D virtual reality techniques, which have been proved to increase the interactivities and enhance the consumers' shopping experience, have been considered a prominent marketing tool and an essential part in developing e-commerce websites. (Fiore & Jin, 2003; Chittaro & Ranon, 2002).

First of all, a 3D virtual reality (VR) e-commerce website resembles the real world shopping environment so that the customers may feel more comfortable while interacting with it. Besides, by providing a more visually attractive shopping interface, it is more likely to trigger the positive feelings and engage the consumers emotionally with the website. Therefore, in the proposed e-commerce interface, a virtual reality living room where the consumers can interact with the furniture products is facilitated in order

Figure 3-4: The process of one of our 3D chair model

to provide the customers with a pleasant shopping environment. Figure 3-4 shows the process of one of our 3D chair model (the first from the right is the model users will see in our VR living room Figure 3-4).

The website is at http://www.vr-solution.com/ with the open source database - MySQL. For the coding languages, we mainly use php and javascript to communicate with the databases and the functions of virtual reality plug-in. The website pages layout, which is shown in Figure 3-5, contains two panels. The left panel contains a virtual reality living room where the users can try the furniture products shown in the right panel.

The users are allowed to choose any kinds of furniture (sofa, chair or table) and see how their selected piece of furniture or combinations fit in the VR living room the combination of all of them to see. The VR living room is designed with EON software and is embedded into the top of left panel. Currently, the file size of this 3D living room plus all the furniture models is about 11MB currently. Users who have never visit our website need to download (our website will direct the users to EON website) and install the plug-in to interact with the VR living room and the furniture models (after downloading, users will need to reboot their machines). Then, users might need to

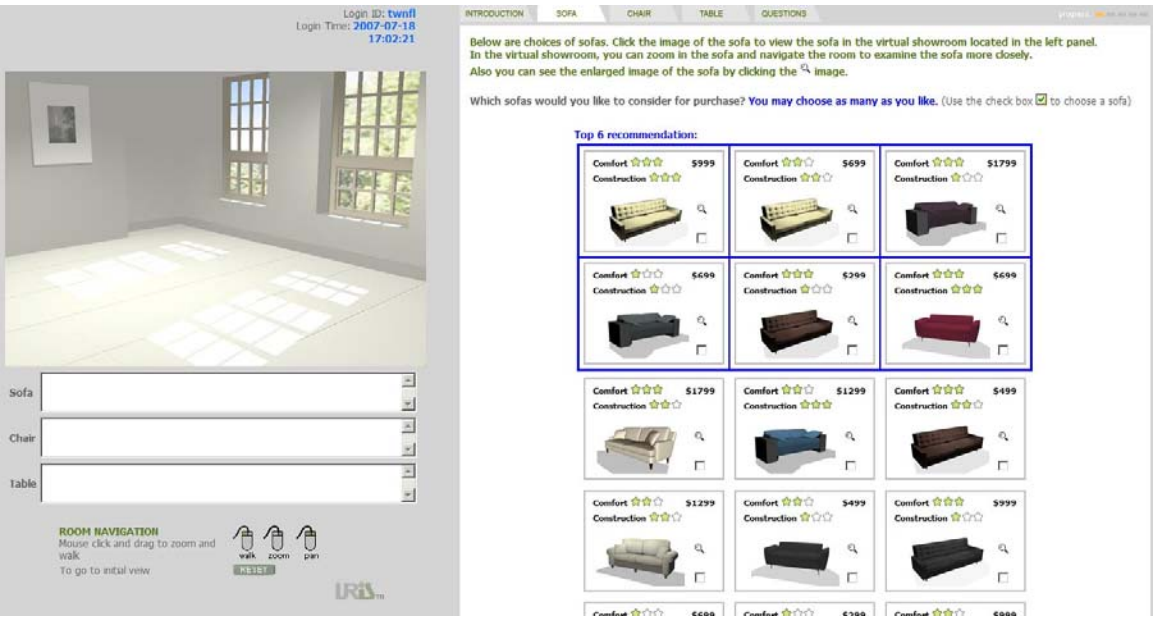

Figure 3-5: Layout of our shopping interface

download the file which is about 11MB. Accordingly, this process might take a while depending on the speed of the Internet and the quality of users' machines.

While interacting with the 3D objects and VR features in this system, the users are provided with additional product details listed below the plug-in living room. For example, in Figure 3-6, one sofa is selected by the user and put into the VR living room, where the additional product information is showing right below. This function applies to every item and every product category.

In the right panel, we provide a list of products with a personalized presentation order that is generated and sorted by our ranking method based on each customer's profile. The top 6 results, varying with the interest of each individual, will be highlighted as our system's recommendations.

Figure 3-6: The left panel is showing the living room with EON plug-in embedded

Figure 3-7 shows part of the sofa list. The consumers can see the pictures, prices, comfort level and construction of all the sofa products and decide which items interest them most. Our shopping website allows customers to select as many items as they want while they are viewing each group of product collection.

Figure 3-7: The right panel is showing the list of products which are sofas in this page

After completing selection of all the product categories, a combination page (see Figure 3-8) will display all the products they have chosen including sofas, chairs, and tables. Then they will have to choose exact only one from sofas, chairs, and tables to be their final decision. During this process, the users are allowed to move back and forward until they make up their mind.

Figure 3-8: The combination page lists the entire favorites have been chosen by the users

Chapter 4

User Profiling

It has long been accepted that there are strong relationships between users' demographic information and their selected products (Vozalis and Margaritis 2004). Such correlations can be defined and represented by determining the similarity of two sets of data, in this case the similarity between demographic profiles and selected products. The similarity is calculated by dot-product of vectors. The details about the Vector-Model method will be discussed in Chapter 5.

In the recommendation system, developing a user profiling system is the first step. We adopt the concepts from demographic filtering and clustering techniques to initially define consumers' characteristics. Demographic and background information, commonly used in marketing segmentation, enables the system to group users with the same personal attributes. Then, the patterns of shopping behavior and final decisions within a group will be analyzed and put into user profiles (Shardanand & Maes, 1995). Through this approach, we can not only find out how the background information relates to a user's purchase decision but also discover the connections of the products purchased.

In the beginning of the project, we conducted a pretest survey (281 samples) to

study the correlation of users' decisions and their personal information (see Section 4.1). Then, after analyzing, we cluster the users based primarily on their demographic information into several groups and levels (see Section 4.2). We also design a profile filtering algorithm to find the most suitable profile for every future incoming customer (see Section 4.3).

4.1 Pretest data

Most of the pretest data were collected from participants at the University of Missouri and the University of Florida. By the time we started analyzing the pretest data, we had 281 valid samples. In this project, a sample is considered valid if the participant had finished all the questionnaires and completed the entire shopping process by submitting their final decisions.

Because the pretest data were mostly provided by college students, the users' demographic profiles were predominantly filled by 18 to 22-year-old college students. In this case, because of the large number of user profiles in the subset "18 to 22 years old" "college students", other subsets in *Age* and *Profession* seem trivial and cannot be used for comparison in the database. Therefore, the two demographic categories, *Age* and *Profession*, were discarded in this project. However, this problem would probably not occur in a real-world situation where a dominant subset should not exist as demographics would be spread across a variety of users.

4.2 Profiles Making

After ruling out the data in *Age* and *Profession*, there are 5 categories used for demographic profiles in this project: *Gender*, *Major*, *Marital Status*, *Ethnicity*, and *Annual Family Income*. All the users were grouped into subsets under each category. All the subsets and the number of samples are shown in Table 4-1. In this project, if a subset contains less than 9 users, it will be considered insignificant because the results are

Table 4-1: Demographic information of all valid samples (N=281)

Demo. Categories	Subset	Samples	Ratio
Gender	Male	140	50%
	Female	141	50%
Major	Business and Finance	138	49%
	Design and Construction	48	17%
	Engineering Sciences	37	13%
	Journalism and Communications	29	10%
Marital Status	Single	255	91%
	Unmarried couple	23	8%
	Married	3	1%
Ethnicity	White	199	71%
	African American	15	5%
	Asian or Pacific Islander	30	11%
	Hispanic	28	10%
	Native American	0	0%
	Other	9	3%
Annual Family Income	under $15,000	46	16%
	$15,000-$24,999	11	4%
	$25,000-$49,999	20	7%
	$50,000-$74,999	29	10%
	$75,000-$99,999	23	8%
	$100,000 or more	93	33%
	Don't know	59	21%

not sufficient to make effective predictions. Thus, two subsets, *Married* under *Marital Status* and *Native American* under *Ethnicity* were neglected.

4.2.1 Profiles - Level 1

As seen in Table 4-1, there are 20 valid demographic subsets in the database. When filtering the data, these individual subsets are defined as "Level 1" profiles since the users are grouped based on one certain type of demographic information. In Level 1, all the data in each subset stand alone, rather than combined with those in another.

However, the recommendation for any incoming customer based on only one kind of personal data would be too generalized. Thus, we attempt to cluster the users based on combinations of personal data in the demographic profiles (5 categories in the current project) to enhance the quality of predictions and recommendations generated by the systems. These profiles are defined as Level M, where M is the number of types of demographic information. For instance, users that share information in three demographic categories, e.g. "Female" and "Single" with income "Under USD 15,000" would be in Level 3.

4.2.2 Profiles - Level 2

It is obvious that data in Level I cannot offer reliable information because of the broad range of the population. Predictions made based solely on a user's *Gender*, *Major*, *Marital Status*, *Ethnicity* or *Income* are too generalized. Therefore, it is necessary to "upgrade" the larger group of "Level 1" clusters into "Level 2", where two demographic

categories are considered.

After re-clustering, there are 62 groups with two demographic characteristics matched. In Table 4-2, the top 30 combinations are listed.

Table 4-2: Level 2 Profiles, sort by N(P*)

P*	N(P*)	Gender	Ethnicity	Marital	Major	Income
2-1	179		White	Single		
2-2	129	Female		Single		
2-3	126	Male		Single		
2-4	126			Single	Business & Finance	
2-5	105	Female	White			
2-6	102		White		Business & Finance	
2-7	94	Male	White			
2-8	83			Single		$100,000 or more
2-9	78	Male			Business & Finance	
2-10	75		White			$100,000 or more
2-11	60	Female			Business & Finance	
2-12	55			Single		Don't know
2-13	54				Business & Finance	$100,000 or more
2-14	49	Female				$100,000 or more
2-15	44	Male				$100,000 or more
2-16	44			Single		under $15,000
2-17	42			Single	Design & Construction	
2-18	40	White				Don't know
2-19	38		White		Design & Construction	
2-20	34			Single	Engineering Science	
2-21	31	Female			Design & Construction	
2-22	30	Female				Don't know
2-23	30		White			under $15,000
2-24	29	Male				Don't know
2-25	28			Single		$50,000-$74,999
2-26	28				Business & Finance	Don't know
2-27	28	Male			Engineering Science	
2-28	27		Asian*	Single		
2-29	27			Single	Journalism & Communication	
2-30	26		Hispanic	Single		

(P* = Profile, Asian* = Asian or Pacific Islander)

4.2.3 Profiles - Level 3+

Unfortunately in the current project, Level 3, Level 4, and Level 5 are unachievable due to the small size of the sample population, namely there were not enough individuals who shared three or more pieces of demographic information. During the user profiling process, the more demographic categories considered, the more personalized and detailed the user profiles can be. In addition, as we only recruited college students as experimental participants rather than a sample pool resembling the real world, there were not enough samples for some subsets when we started analyzing the pretest data.

It is important to point out that this survey website is an ongoing project and is still running to collect data. Hopefully, we will have “Level 3”, “Level 4”, and even “Level 5” in the future, once we have collected a larger and more representative sample. With a larger, more representative sample, we might be able to expand our profiling, by including the two categories *Age* and *Profession* and by adding more demographic or lifestyle categories, once data from participants outside the campus is gathered.

4.3 Profile Filter

As soon as a new user registers, his/her information is categorized into different subsets under the main demographic categories. Then, the profile filtering system would automatically scan the profiles starting from Level M, where the most various combinations of subsets are contained, until it finds a match profile. The scheme of our profile filter is illustrated in Figure 4-1.

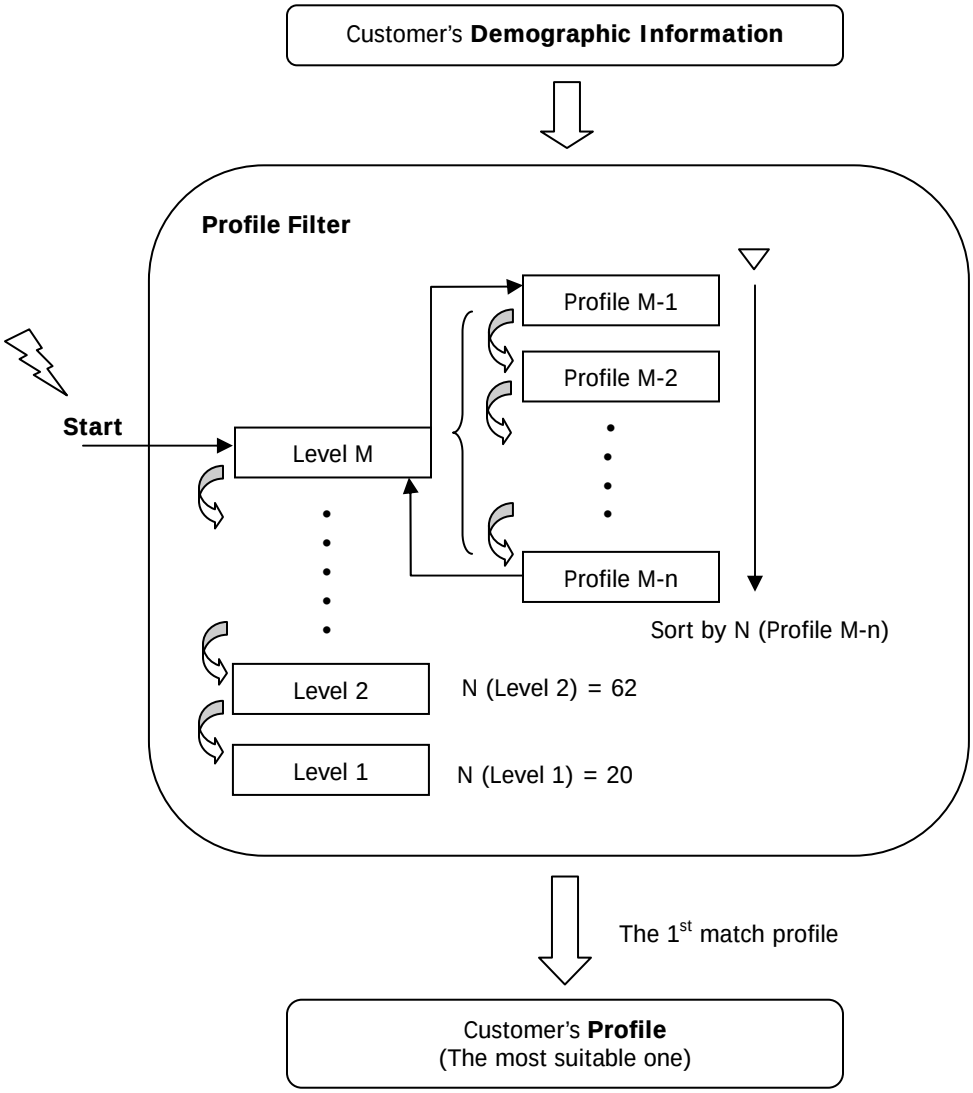

Figure 4-1: Scheme of User Profiling

Our profiling system consists of the following terms:

Profile: Each profile includes information about a group of users that share the same demographic background as well as their shopping behaviors and final decisions.

Level M: The users in this set of profiles have M demographic data in common. The

maximum number for M is the total number of demographic categories..

N (Level M): The number of profiles in Level M.

Profile M-n: The nth profile in Level M.

N (Profile M-n): The number of samples from pretest data in Profile M-n.

If we take Level 2 for an example, N (Level 2) = 62 can be described as:

1) There are 62 profiles in Level 2.
2) The users in Level 2 have two demographic categories in common.
3) Table 4-2 shows the top 30 profiles in Level 2, ranked according to the number of customers in the profiles.
4) The first row in Table 4-2:
    i. We call it Profile 2-1 because it has the most number of customers in this level.
    ii. N (Profile 2-1) = 179 means number of customers in Profile 2-1 is 179.
    iii. The 179 customers in Profile 2-1 share two pieces of demographic information, namely *White* and *Single*.

Chapter 5

Hybrid Ranking Method

The previous chapter introduces the design of our filtering system in categorizing the customers into different user profiles, which stores “the user's interests (positive as well as negative) in specific items” (Shardanand & Maes, 1995) as well as their demographic information. However, since each movement or choice a user makes is a significant indicator of his/her interests or needs, an adaptive recommendation system should be able to adjust itself dynamically while the user is interacting or navigating through the website. Therefore, we integrate the demographic filtering approach with other two techniques, the “Vector-Space Model” in information retrieval (Baeza-Yates & Ribeiro-Neto, 1999) and “Association Rules” in data mining (Agrawal, Imielinski, & Swami, 1993; Agrawal & Srikant, 1994), into a hybrid ranking system. The Vector-Space Model and Association Rules are highly effective tools for analyzing the relationships between/among users’ navigation and purchasing behaviors.

5.1 Vector-Space Model

Information retrieval technology has become a compelling topic as people attempt to

overcome the information overload problem. One of the information retrieval techniques is the Vector-Space Model, which uses vectors to present two sets of data, such as queries and electronic documents. The level of their correlation depends on the similarity of the two vectors. For instance, given a query, the most useful document related to the query would be the one with the smallest similarity in the vector-space model. The following paragraph introduces the Vector-Space Model as an information retrieval tool for managing a document database. How this approach is applied in the current e-commerce recommendation system is thoroughly discussed in Section 5.1.1 to 5.1.5.

In their book, Baeza-Yates and Ribeiro-Neto (1999) state that all documents can be described by a set of keywords. Thus, when retrieving information, a system can find a target document immediately by comparing the keyword set rather than parsing the whole text. However, while the keyword set presents a “logical view” that describes the document, it cannot represent the entire document. Therefore, the index terms, the minimum set of words from the full text after elimination of words such as connectives, articles and adjectives, are the most efficient way to identify a document. In this project, the index term is defined as each product feature that identifies the furniture item.

5.1.1 Definition

Originally, the vector-space model was designed as a tool for search engines. Since our system is a shopping-oriented interface, we have modified its semantic definition to fit our e-commerce site. First of all, we treat every product as a document and define the user profiles (denoted as Profile M-n) that have been generated by our profiling filtering method (see Chapter 4) as the user query. Hence, for any customer, provided with his/her

profile, our system can find the closest products using this method.

Let $w_{i,p}$ be the weight associated with the pair $[k_i, p]$, where $w_{i,p} \geq 0$. Each user profile's vector $\vec{p}$ is defined as $\vec{p} = (w_{1,p}, w_{2,p}, \dots, w_{t,p})$ where *t* is the total number of index terms (see Section 5.1.2). For each product d_j, its vector is defined as $\overrightarrow{d_j} = (w_{i,j}, w_{2,j}, \dots, w_{t,j})$. Therefore, a profile P and a product d_j can be represented as *t*-dimensional vectors as shown in Figure 5-1.

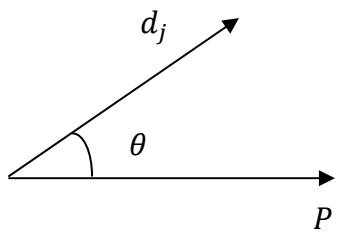

Figure 5-1: *t*-dimensional vectors and the angle θ

The vector model is designed to find the correlation of vectors by evaluating the degree of similarity (Baeza-Yates & Ribeiro-Neto, 1999; Lee, Chuang, & Seamons 1997). In this project, we assume that

- ➔ relevant products are similar to each other
- ➔ irrelevant products are not similar to the relevant products
- ➔ the most relevant product for the profile is the one that is the closest to the profile

A common similarity measure, known as the *cosine measure*, is adopted to find the cosine of angle θ between vectors which is defined as:

$$sim(d_j, p) = \frac{\vec{d_j} \bullet \vec{p}}{|\vec{d_j}| \times |\vec{p}|}$$

$$= \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,p}}{\sqrt{\sum_{i=1}^{t} {w_{i,j}}^2} \times \sqrt{\sum_{i=1}^{t} {w_{i,p}}^2}} \tag{5.1}$$

We also adopt a very common approach, the so-called $tf \times idf$ method in which the weight of a term is determined by two factors: how often the index term k_i occurs in each product/profile (the term frequency $tf_{i,j}$ which is shown in Equation 5.2) and how often it occurs in the whole product category or user profile (the product/profile frequency df_j). In Equation 5.2, $max(freq_{i,j})$ is the maximum number of overall $freq_{i,j}$ which is the frequency of index term k_i found in the product d_j. For example, if the index term k_i does not occur in the product d_j, then its $tf_{i,j}$ equals to 0.

$$tf_{i,j} = \frac{freq_{i,j}}{max(freq_{i,j})} \tag{5.2}$$

Further, let N be the total number of products/profiles. Then the equation of idf_i which stands for inverse product frequency would be Equation 5.3.

$$idf_j = log \frac{N}{df_j} \tag{5.3}$$

The term-weighting formula is defined as Equation 5.4.

$$w_{i,j} = tf_{i,j} \times idf_j$$

$$= tf_{i,j} \times log \frac{N}{df_j} \tag{5.4}$$

Several variations of Equation 5.2 have been proposed in prior works. In our system, we adopt the method which is suggested by Salton and Buckley(1988). It is shown in Equation 5.5.

$$tf_{i,j} = 0.5 + \frac{0.5\, freq_{i,j}}{max(freq_{i,j})} \tag{5.5}$$

Then the new term-weighting method derived from Equation 5.4 is shown in Equation 5.6.

$$\begin{aligned} w_{i,j} &= tf_{i,j} \times idf_j \\ &= \left(0.5 + \frac{0.5\, freq_{i,j}}{max(freq_{i,j})}\right) \times log\, \frac{N}{df_j} \end{aligned} \tag{5.6}$$

5.1.2 Index Terms

In order to progressively obtain and predict personal interests, we cluster products into different groups including sofas, chairs, and tables with each webpage displaying exactly one product category respectively. The first product page shows sofas only; the second page shows the chairs, and the third page displays the tables. In this system design, these three product categories share the same equations.

For example, to recommend our first webpage which shows first product category (sofas) to a customer whose user profile is *Profile M-n*, we first need to define the number of index terms (*t*) to calculate the *t*-dimensional vectors for *Profile M-n* and all sofas.

Table 5-1: The attributes of a sofa

Category		Index Terms (k_i)
ID	sofa4	
Name	Casual_S1_M2	sofa_type_1
Price	1799	sofa_price_3
Construction	3	sofa_construction_3
Comfort	3	sofa_comfort_3
Body	Fabric-Blue	
Frame	Walnut	
Textile	Fabric	
Color	Vivid Color	sofa_color_8
Design	Medium Wood	sofa_design_3

Each product is assigned with an index card which contains all the information about that item, including product pictures, product features, its index terms and other details. Table 5-1 is an example showing the index card of one sofa product. The index cards are for internal use only. The customers are only provided with the product picture and the feature information marked in color.

Due to the various features of products (see Section 3.3), we define each sofa and chair with 6 index terms (type, price, construction, comfort, color and design) and each table with 5 index terms (type, price, construction, maintenance and design) according to the results in the pretests. In total, there are 24 index terms (t) in the sofa category, 22 in the chair category and 17 in the table category. Table 5-2 shows the complete list of all index terms for these three categories. The similarity $sim(d_j, p)$ between *Profile M-n* and all the sofas can then be calculated with 24-dimensional vectors, with 22-dimensional vectors for chairs, and with 17-dimensonal vectors for tables. The concept is illustrated in Figure 5-2.

Table 5-2: The list of all index terms in three product categories

	Sofas	Chairs	Tables
Index Terms (k_i)	sofa_type_1	chair_type_1	table_type_1
	sofa_type_2	chair_type_2	table_type_2
	sofa_type_3	chair_type_3	table_type_3
	sofa_price_1	chair_price_1	table_price_1
	sofa_price_2	chair_price_2	table_price_2
	sofa_price_3	chair_price_3	table_price_3
	sofa_construction_1	chair_construction_1	table_construction_1
	sofa_construction_2	chair_construction_2	table_construction_2
	sofa_construction_3	chair_construction_3	table_construction_3
	sofa_comfort_1	chair_comfort_1	table_maintenance_1
	sofa_comfort_2	chair_comfort_2	table_maintenance_2
	sofa_comfort_3	chair_comfort_3	table_maintenance_3
	sofa_color_1	chair_color_1	table_design_1
	sofa_color_2	chair_color_2	table_design_2
	sofa_color_3	chair_color_3	table_design_3
	sofa_color_4	chair_color_4	table_design_4
	sofa_color_5	chair_color_5	table_design_5
	sofa_color_6	chair_color_6	
	sofa_color_7	chair_color_7	
	sofa_color_8	chair_design_1	
	sofa_design_1	chair_design_2	
	sofa_design_2	chair_design_3	
	sofa_design_3		
	sofa_design_4		

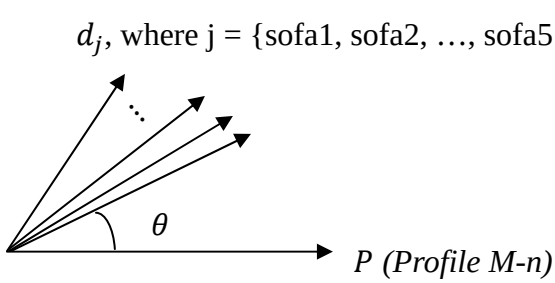

Figure 5-2: 24-dimensional vectors and the angle θ for all sofas and the *Profile M-n*

5.1.3 Product Vector $\overrightarrow{d_j}$

In Equation 5.5, whenever an index term k_i occurs in the product d_j, $freq_{i,j}$ would be always equal to 1, because it occurs exactly once in any product d_j. Also, $max(freq_{i,j})$ would always be equal to 1 since $freq_{i,j}$ is either equal to 1 or 0. Consequently, for any index terms (k_i) that occur in any product category d_j (sofas, chairs, or tables), $tf_{i,j}$ would be always equal to 1. This is demonstrated in the following equation.

$$tf_{i,j} = 0.5 + \frac{0.5\, freq_{i,j}}{max(freq_{i,j})} = 0.5 + \frac{0.5 \times 1}{1} = 1 \tag{5.7}$$

On the other hand, $tf_{i,j}$ is simply equal to 0 if the index term k_i never occurs in the product d_j.

The product frequency df_j is defined as how often the index term k_i occurs in the whole products collection d_j. For example, in the first row in Table 5-3, $df_j = 18$ means there are 18 sofas in the sofa category d_j (where j=sofa1~sofa51) that are "Type – Casual" (index term k_i = "sofa_type_1"), 16 that are "Type - Modern", and the rest are "Type - Traditional". Therefore, we can always get the exact size of the product category by adding up df_j from each feature category (which is 51 for sofas). For example, the summing df_j when the feature category is "design" (sofa_design1~sofa_design4) is equal to 6+8+24+13=51. In addition, since we have d_j for each index term k_i and the N is 51 in the sofa category, we can compute $idf_j (= log \frac{N}{df_j})$. The results are shown in Table 5-2.

Table 5-3: The Index Terms (k_i), df_i and idf_j of sofa category for d_j

Feature Category		Index Terms (k_i)	df_i	idf_j
Type	Casual	sofa_type_1	18	0.452298
	Modern	sofa_type_2	16	0.50345
	Traditional	sofa_type_3	17	0.477121
price	Cheap	sofa_price_1	12	0.628389
	Fine	sofa_price_2	18	0.452298
	Expensive	sofa_price_3	21	0.385351
construction	Fair	sofa_construction_1	17	0.477121
	Good	sofa_construction_2	22	0.365148
	Superior	sofa_construction_3	12	0.628389
comfort	Fair	sofa_comfort_1	16	0.50345
	Good	sofa_comfort_2	15	0.531479
	Superior	sofa_comfort_3	20	0.40654
color	Dark Color	sofa_color_1	4	1.10551
	Dark Neutral	sofa_color_2	13	0.593627
	Green	sofa_color_3	4	1.10551
	Light Neutral	sofa_color_4	2	1.40654
	Light Color	sofa_color_5	3	1.23045
	Medium Color	sofa_color_6	2	1.40654
	Pattern	sofa_color_7	11	0.666178
	Vivid Color	sofa_color_8	12	0.628389
design	Dark Wood	sofa_design_1	6	0.929419
	Light Wood	sofa_design_2	8	0.80448
	Medium Wood	sofa_design_3	24	0.327359
	Metal	sofa_design_4	13	0.593627

Using Equation 5.8, 24-dimensional vectors for all sofas can then be calculated.

For the purpose of efficiency, the results shown in Table 5-2 were computed offline and stored in our database since all these results are static to all profiles until we modify the products design and attributes in the future.

$$\overrightarrow{d_j} = [w_{k_1,j} \quad \cdots \quad w_{k_{i-1},j}, w_{k_i,j}]$$

$$where\ w_{i,j} = tf_{i,j} \times idf_j = \begin{cases} idf_j\ , & if\ k_i \in d_j \\ 0\ , & otherwise \end{cases} \tag{5.8}$$

5.1.4 User Profile Vector $\vec{p}$

In this section, we use *Profile 1-5* (the group of users who have the same major – Business and Finance) as an example to demonstrate the process in generating *t*-dimensional vectors for all profiles. By the rules mentioned in Chapter 4, the users in this profile share only one condition in common, their major – Business and Finance. Thus, their profile belongs to Level 1. In addition, because the number of users in this profile is 138, which is the 5th largest in Level 1, this group is denoted as *Profile 1-5* with *N(Profile 1-5)=138*.

The vectors of *Profiles 1-5* and sofas will be 24-dimensional, as per the discussion in the previous section. One thing we have to mention carefully is that df_i, which stood for product frequency in a product category in previous section, will stand for profile frequency here. Profile frequency df_i means how often the index term k_i occurs among all the final choices (see Table 3-2) from the users in this profile. Table 5-4 shows the results of profile frequency df_i with all the index terms (k_i) in *Profiles 1-5*. For each category, the sum of df_i is equal to the size of pretest samples in this profile. For example, the df_i in construction category, 37+57+44 = N(Profile 1-5) =138. Also, idf_j can be calculated in Equation 5.3 where N equals to 138.

To compute $tf_{i,p}$ for this profile, in Equation 5.5, $freq_{i,p}$ is simply equal to the profile frequency df_i. Therefore, looking for $max(freq_{i,p})$ is equal to finding the maximum of all df_i which is 93 from *Profile 1-5*. For example, the $tf_{i,p} = 0.5 + (0.5\ freq_{i,p}/\ max(freq_{i,p})) = 0.5 + (0.5 \times 22/93) = 0.61828$, where index term ($k_i$) is "sofa_comfort_1".

Table 5-4: The Index Terms (k_i), and df_i of sofas collection in *Profile 1-5*

Feature Category		Index Terms (k_i)	df_i
Type	Casual	sofa_type_1	61
	Modern	sofa_type_2	71
	Traditional	sofa_type_3	6
price	Cheap	sofa_price_1	37
	Fine	sofa_price_2	55
	Expensive	sofa_price_3	46
construction	Fair	sofa_construction_1	37
	Good	sofa_construction_2	57
	Superior	sofa_construction_3	44
comfort	Fair	sofa_comfort_1	22
	Good	sofa_comfort_2	23
	Superior	sofa_comfort_3	93
color	Dark Color	sofa_color_1	14
	Dark Neutral	sofa_color_2	60
	Green	sofa_color_3	4
	Light Neutral	sofa_color_4	10
	Light Color	sofa_color_5	0
	Medium Color	sofa_color_6	10
	Pattern	sofa_color_7	4
	Vivid Color	sofa_color_8	36
design	Dark Wood	sofa_design_1	21
	Light Wood	sofa_design_2	0
	Medium Wood	sofa_design_3	61
	Metal	sofa_design_4	56

According to the idf_j and $tf_{i,p}$ just mentioned, the term-weighting method for user profiles is shown in Equation 5.9.

$$\vec{p} = [w_{k_1,p} \quad \cdots \quad w_{k_{i-1},p}, w_{k_i,p}]$$

$$where\ w_{i,p} = tf_{i,p} \times idf_j = \begin{cases} tf_{i,p} \times idf_j\ , & if\ k_i \in p \\ 0 & ,\ otherwise \end{cases} \tag{5.9}$$

5.1.5 Similarity of d_j and p

Currently, fifty-one 24-dimensional vectors have been generated in the sofa category $\overrightarrow{d_j}$ (see Section 5.1.3) and profile $\vec{p}$ (see Section 5.1.4) for sofa recommendation. The similarity measure we are using is the *cosine measure* (Equation 5-1), meaning that he more similar two 24-dimensional vectors are, the higher the value returned from the cosine measure. The highest score is 1 and occurs when the angle θ between the two vectors is 0. Table 5-3 shows the top 10 results of similarity between sofa collection $\overrightarrow{d_j}$ and Profile 1-5 $\vec{p}$.

Table 5-5: The top 10 ranking results of sofas for Profile 1-5

Rank.	Sofa ID	Index Terms			Score
1	sofa37	sofa_type_2 sofa_comfort_3	sofa_price_2 sofa_color_6	sofa_construction_3 sofa_design_4	0.522742
2	sofa20	sofa_type_2 sofa_comfort_2	sofa_price_2 sofa_color_6	sofa_construction_2 sofa_design_4	0.492883
3	sofa6	sofa_type_2 sofa_comfort_3	sofa_price_3 sofa_color_1	sofa_construction_1 sofa_design_1	0.480126
4	sofa29	sofa_type_2 sofa_comfort_1	sofa_price_2 sofa_color_1	sofa_construction_1 sofa_design_1	0.47925
5	sofa14	sofa_type_2 sofa_comfort_3	sofa_price_1 sofa_color_2	sofa_construction_1 sofa_design_4	0.455206
6	sofa7	sofa_type_2 sofa_comfort_3	sofa_price_2 sofa_color_8	sofa_construction_3 sofa_design_4	0.453516
7	sofa18	sofa_type_1 sofa_comfort_3	sofa_price_3 sofa_color_4	sofa_construction_2 sofa_design_3	0.452478
8	sofa10	sofa_type_2 sofa_comfort_2	sofa_price_3 sofa_color_8	sofa_construction_3 sofa_design_1	0.452206
9	sofa25	sofa_type_2 sofa_comfort_3	sofa_price_1 sofa_color_2	sofa_construction_2 sofa_design_4	0.450729
10	sofa49	sofa_type_1 sofa_comfort_2	sofa_price_3 sofa_color_4	sofa_construction_2 sofa_design_3	0.445869

5.1.6 User Relevance Feedback

Relevance feedback is the most popular query reformulation strategy. Presented with a ranked list of products d_j, the user's real-time interactions not only give feedback on the products that have been recommended but also provide relevant information to generate future predictions. There are two ways to increase the accuracy of our ranked product list: expand the set of index terms in Profile p, and adjust the term weights in Profile p. The expected effect is that the new Profile p' will be moved towards the relevant products and away from the non-relevant products. Early studies, such as Smart System [6], have shown significant improvement for small test collections when relevance feedback is used (Baeza-Yates & Ribeiro-Neto, 1999).

The definitions of some additional terminology used regarding the processing of a given Profile p are as follows:

$\vec{\mathrm{q}}$: original query

$\overrightarrow{\mathrm{q_m}}$: modified query

$\vec{\mathrm{p}}$: original profile (for sofas and chairs)

$\overrightarrow{\mathrm{p'}}$: modified profile (for chairs and tables)

D_r : a set of index terms of products that the system is going to recommend based on the user's dynamic feedback on the products that have been suggested.

$|D_r|$: the size (number) of D_r

$\overrightarrow{d_j}$: t-dimensional vectors representing D_r (t: the number of index terms of a product collection)

α, β, γ: tuning constants

There are three classic ways to calculate the modified query. The first is Ide_Dec_Hi, which is the one adopted in the current system. The original Ide_Dec_Hi equation is shown below:

$$\overrightarrow{q_m} = \alpha\vec{q} + \beta \sum_{\forall \overrightarrow{d_j} \in D_r} \overrightarrow{d_j} - \gamma \, max_{non-relevant}\left(\overrightarrow{d_j}\right) \tag{5.10}$$

The denotation $max_{non-relevant}$ is a reference to the highest ranked non-relevant document (product). However, since the current shopping interface does not provide the function for the consumers to simply vote on products they are not interested in at all (non-relevant), this parameter is not considered. Therefore, the adapted equation is:

$$\overrightarrow{p'} = \alpha\vec{p} + \beta \sum_{\forall \overrightarrow{d_j} \in D_r} \overrightarrow{d_j} \tag{5.11}$$

In the original formulation, Rochio [7] fixed $\alpha = 1$ and Ide [7] fixed $\alpha = \beta = \gamma = 1$. The main advantages of the three relevance feedback techniques mentioned above are their simplicity and their strong ability to lead to significantly improved results. The simplicity is due to the fact that the modified term weights are computed directly from all products. The improvement in results is observed experimentally and is due to the fact that the modified profile vector does reflect a portion of the intended profile semantics.

Nevertheless, on this e-commerce interface, because the consumers make multiple choices of items they are interested in rather than simply voting for the ones they consider “relevant”, whether their choices are relevant to the recommendations are

analyzed and revealed by the system. For this reason, we consider the multiple choices of all products from each product category (sofa, chair and table) consumers made as the set of relevant products D_r defined by consumers.

A major difference between the original User Relevance Feedback and our system is that for User Relevance Feedback (Equation 5.10), after receiving the user's feedback, the modified query $\overrightarrow{q_m}$ is the modification of original query $\vec{q}$; however, in this project, instead of re-weighting the profile $\overrightarrow{p}$ on the sofa list and re-ranking the sofa recommendation list after the consumer has chosen his/her favorite sofas, the system will change the presentation order of the next product category, chair, by weighting the profile $\overrightarrow{p}$ on the chair recommendation list.

For example, assume that a consumer in Profile 1-5 logs into the website. As soon as he/she logs in, the rankings of all the three product categories will be ready but only the sofa lists will be provided. The ranking of all sofas is represented by $\overrightarrow{p_{1-5\ on\ sofa}}$, and the ranking of chair collection is represented by $\overrightarrow{p_{1-5\ on\ chair}}$. At the moment when he/she picks up the favorite sofas and click "Next" onto the chair page, the index terms of these picked sofas will be considered as user relevance feedback. The system will re-rank the original ranking $\overrightarrow{p_{1-5\ on\ chair}}$ by weighting those index terms. So, the $\vec{p}$ in Equation 5.11 is $\overrightarrow{p_{1-5\ on\ chair}}$, and $\overrightarrow{p_{1-5\ on\ chair}}'$ is what the system is looking for. Figure 5-3 shows the scheme to illustrate the above process when a customer in Profile 1-5, denoted as C_{1-5} , visits the e-commerce website.

However, since the three product categories do not share any index terms (see Table 5-2), the system has to rely on another technique, association rules, to determine the relationships between index terms and product categories.

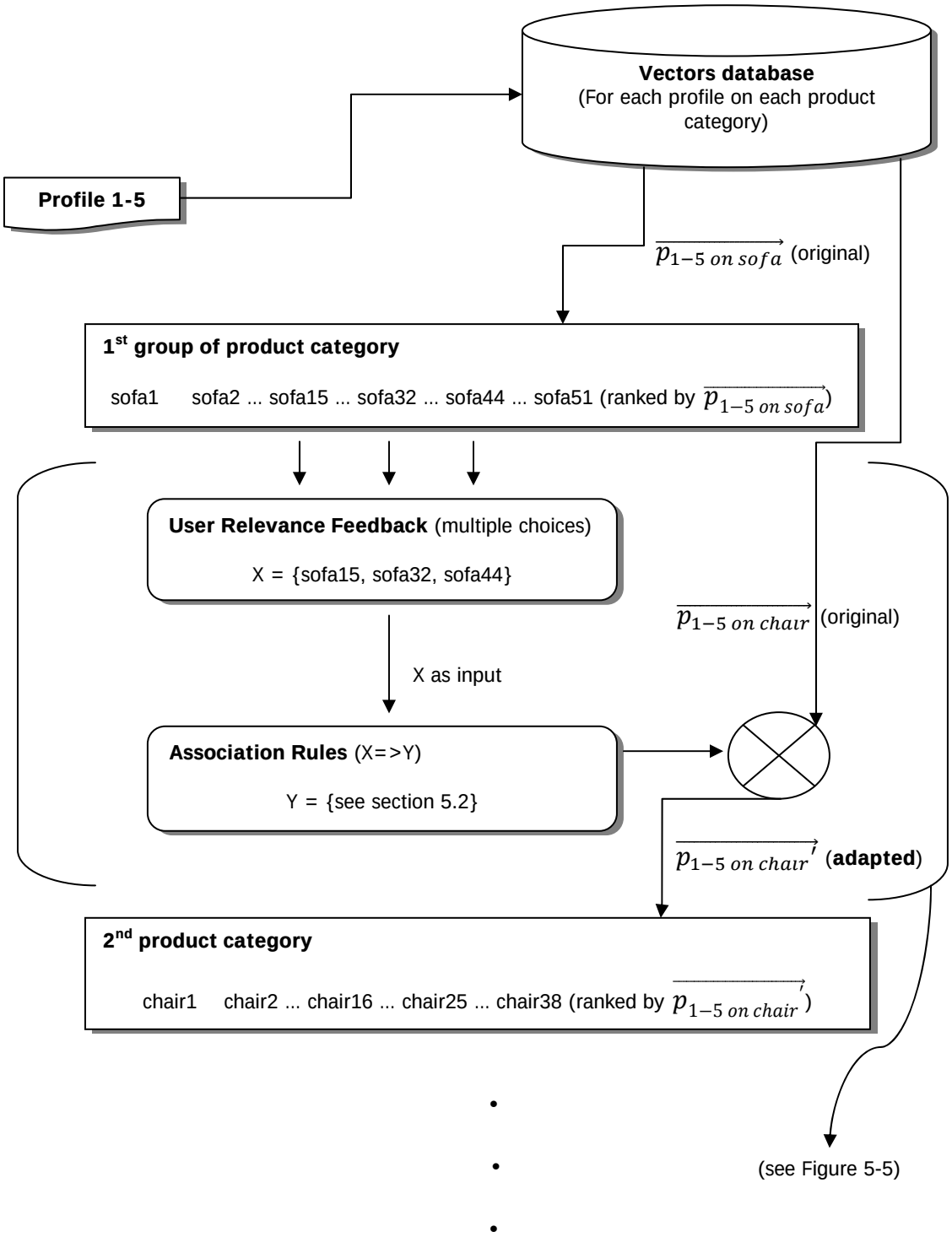

Figure 5-3: A scheme to demonstrate the procedure of generating recommendations on the first two product categories for the customer C_{1-5} (continued, see Figure 5-5 for the Association Rules in detail)

5.2 Association Rules

Association rules (Agrawal et al., 1993; Agrawal & Srikant, 1994) have been commonly used in data mining technology. “Data mining is, in some ways, an extension of statistics, with a few artificial intelligence and machine learning twists thrown in. (Thearling 1995)” It is used to discover some hidden but meaningful or interesting information by analyzing a large amount of data stored in the database.

In data mining (Dunham 2003), Association Rules is an approach that analyzes the characteristics between sets of data in order to find out those similar patterns or relationships that have been highly relied on in merchandising. By analyzing the patterns of the customers’ purchase behaviors, association rules are able to reveal the preferences of products, correlations between each purchase and further predict the consumers’ choices. For example, when one product is often purchased with another one, there is an association. This technique is frequently used by retail stores to assist in areas such as marketing, advertising, product promotion, and product placement.

However, the traditional store can only find out the relationships from previous behaviors. By applying association rule technology to e-commerce interfaces, the adaptive system can not only find out a user’s preferences based on the shopping history, but also adjust the presentation of products dynamically according to the user’s navigation behaviors.

5.2.1 Definition

“Given a set of items $I = \{I_1, I_2, \dots I_m\}$ and a database of transactions $D = \{t_1, t_2, \dots, t_n\}$

where $t_i = \{I_{i1}, I_{i2}, \ldots, I_{ik}\}$ and $I_{ij} \in I$, an association rule is an implication of the form $X \Rightarrow Y$ where $X, Y \subset I$ are sets of items called itemsets and $X \cap Y = \emptyset$ (Dunham 2003)."

The **support (s)** for an association rule $X \Rightarrow Y$ is the percentage of transactions in the database that contain $X \cup Y$, and the **confidence (α)** for an association rule $X \Rightarrow Y$ is the ratio of the number of transactions that contain $X \cup Y$ to the number of transactions that contain *X*. [5]

The Apriori algorithm is one of the prevailing techniques to find association rules (Agrawal, et al., 1993; Agrawal and Srikant, 1994). ARMiner (Cristofor 2000) is a client-server data mining application specialized in finding association rules. We resort to this software and run experiments with the Apriori algorithm to find the association rules. ARMiner requires a special data format to generate association rules; so, we first write some PHP scripts to massage the data into the correct format for all user profiles (82 in total). Since part of our hybrid ranking method is an on-line process, if there are too many association rules for each profile, the efficiency will drop dramatically. However, for some profiles, the number of association rules is too few, which may negatively influence on our hybrid ranking method as well. Therefore, it is crucial to have a reasonable number of association rules for every profile when designing this system. In order to find out the most appropriate settings, we have experimented on different hypothetical settings of ***support*** and ***confidence***. The results indicate the best performance for ***support=0.4*** and ***confidence=0.7***.

5.2.2 Implementation

Currently, there are only three product categories in our system. Based on the arrangements of product placement on this shopping interface, there are three associations that we are interested in: the choices between “sofa and chair”, “chair and table”, and “sofa and table”, shown in Figure 5-4:

Rule Group 1: $X \Rightarrow Y$, where $X \subset d_j$ (sofa category), and $Y \subset d_j$ (chair category)

Rule Group 2: $X \Rightarrow Y$, where $X \subset d_j$ (chair category), and $Y \subset d_j$ (table category)

Rule Group 3: $X \Rightarrow Y$, where $X \subset d_j$ (sofa category), and $Y \subset d_j$ (table category)

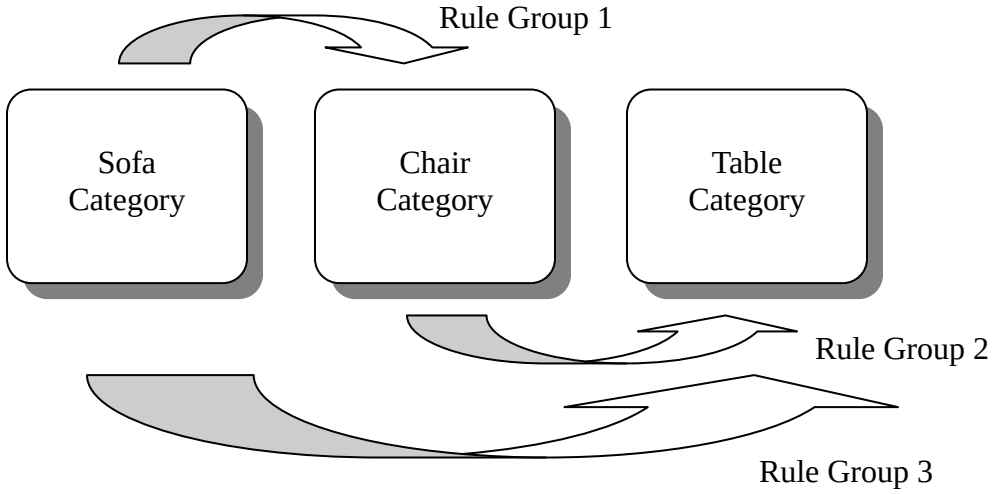

Figure 5-4: Association rules in this adaptive system

After login, the first webpage presented is the personalized product presentation of sofa category that has been ranked by profile. The user is allowed to choose multiple sofas that he/she is interested in. Then he/she can continue to the chair page, choose some chairs, and finally proceed to the table page. Therefore, only the presentation of chairs

and tables need to be re-ranked. This is done by re-weighting technique mentioned in Equation 5.11.

The information for re-ranking the chairs is in Rule Group 1, the =set of association rules that deal with sofa and chair preferences. For the re-ranking of tables, we will need to check on both Rule Group 2 and Rule Group 3.

For example, Table 5-6 shows the results of all the Rule Groups on Profile 1-5 with the parameters ***support=0.4*** and ***confidence=0.7***, and Figure 5-5 shows the scheme to complement Figure 5-3. The association rules stored in our Association Rules Database are the correlations of index terms of each item in the three product categories. Accordingly, we analyze the index terms of the sofas chosen by customer C_{1-5}.

Table 5-6: Association Rules for Profile 1-5 with **support** = 0.4 and **confidence** = 0.7

X	Index Terms of X	Y	Index Terms of Y	support	confidence
Rule Group 1					
Sofa Design - Metal	sofa_design_4	Chair Design - Metal	chair_design_3	0.405797	1
Sofa Design - Medium Wood	sofa_design_3	Chair Design - Medium Wood	chair_design_2	0.442029	1
Sofa Color - Dark Neutral	sofa_color_2	Chair Color - Dark Neutral	chair_color_1	0.434783	1
Sofa Comfort - Superior	sofa_comfort_3	Chair Comfort - Superior	chair_comfort_3	0.521739	0.774194
Sofa Type - Modern	sofa_type_2	Chair Design - Metal	chair_design_3	0.405797	0.788732
Sofa Type - Casual	sofa_type_1	Chair Design - Medium Wood	chair_design_2	0.442029	0.95082
Rule Group 2					
Chair Design - Metal	chair_design_3	Table Design - Metal	table_design_5	0.405797	1
Chair Design - Medium Wood	chair_design_2	Table Design - Medium Wood	table_design_4	0.442029	1
Rule Group 3					
Sofa Design - Metal	sofa_design_4	Table Design - Metal	table_design_5	0.405797	1
Sofa Design - Medium Wood	sofa_design_3	Table Design - Medium Wood	table_design_4	0.442029	1
Sofa Type - Modern	sofa_type_2	Table Design - Metal	table_design_5	0.405797	0.788732
Sofa Type - Casual	sofa_type_1	Table Design - Medium Wood	table_design_4	0.442029	0.95082

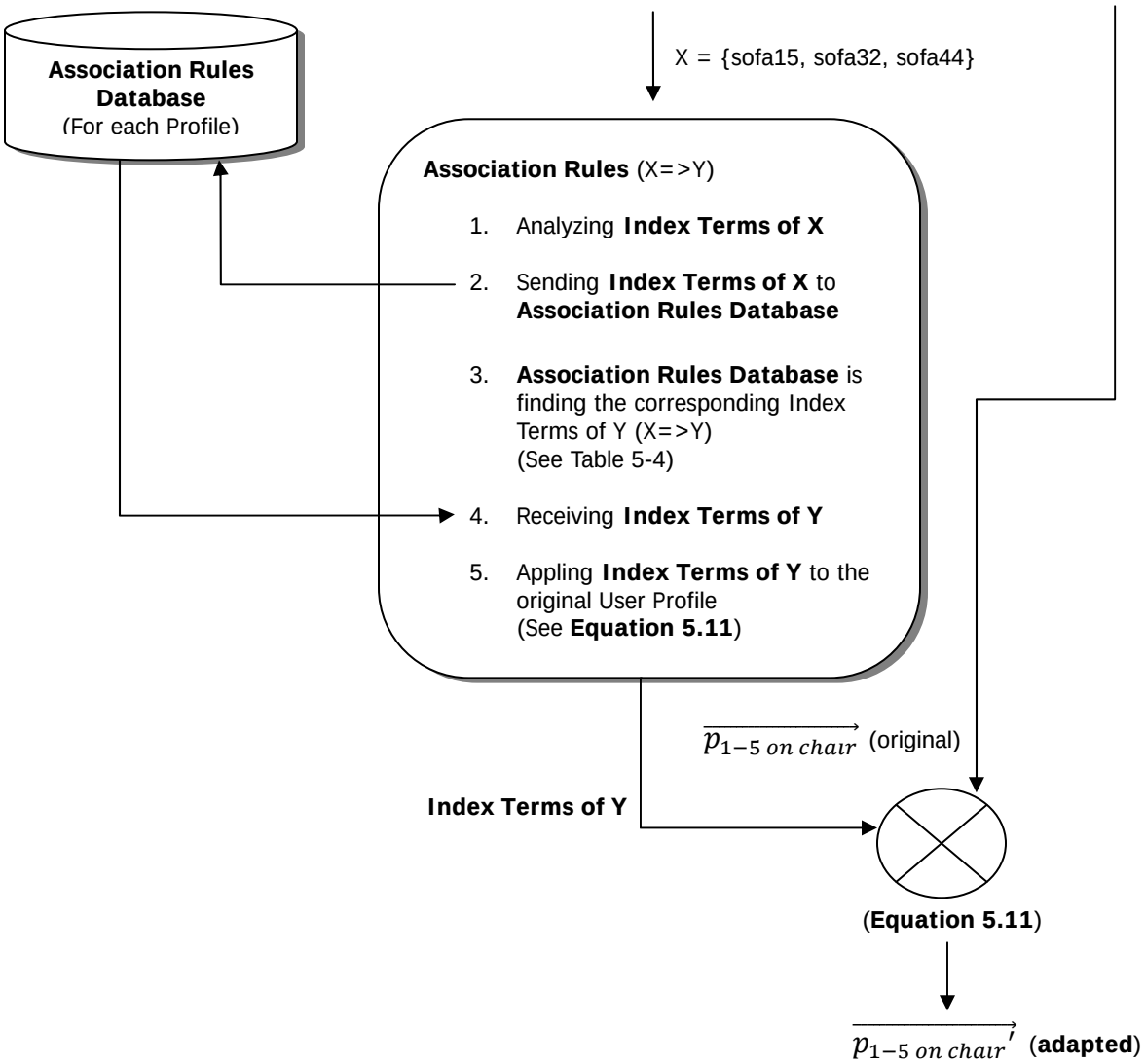

Figure 5-5: A scheme to demonstrate the procedure of generating recommendations on the first two product categories for a customer C_{1-5} (continue Figure 5-3)

These index terms are collectively taken to be the set X and are sent to our Association Rules Database. In the phase from the sofa category X to chair category Y, only Rule Group 1 is applicable. Our Association Rules Database will return the index terms of Y. For instance, if the 3rd rule (3rd row) of Rule Group 1 in Table 5-6 matches the

X, the association rule method is expressing that users who purchase a sofa with color “Dark Neutral” tend to purchase a chair with color “Dark Neutral” at the same time with a 52% probability. Afterwards, for the index terms of chairs which meet Y, we impose more weight on them to cater to the interest of this customer C_{1-5}.

Chapter 6

System Implementation

In Chapters 4 and 5, we have thoroughly discussed the techniques employed, including user profiling, vector-space models, user relevance feedback, and association rules, and the reasons why they were adopted in this system. To illustrate how this system works in practice, this chapter elaborates on the functions of these techniques and how they progressively cater the VR e-commerce website to the user's needs by giving an example using the false user C_{1-5}.

6.1 Procedures

At the moment when user C_{1-5} logs in, the user profiling system (see Chapter 4) begins analyzing his demographic information and determines that Profile 1-5 (Major: Business & Finance) is the most suitable user profile for him. Since the hybrid ranking method (see Section 5.1.1~5.1.5) has already generated the ranking of every profile (20 profiles from Level 1 and 62 from Level 2) on all the product categories during the off-line process, the ranking of sofas, denoted as $\overrightarrow{p_{1-5\ on\ sofa}}$, the ranking of chairs $\overrightarrow{p_{1-5\ on\ chair}}$, and the ranking of table $\overrightarrow{p_{1-5\ on\ table}}$ are ready to be extracted (see Equation 5.9).

The first page customer C_{1-5} exposed to is the sofa page with product recommendation list $\overrightarrow{p_{1-5\ on\ sofa}}$ (see Table 5-5). Customer C_{1-5} chooses some sofas that he is interested in, e.g. “sofa14”, “sofa8”, and “sofa46,” and moves on to the next product category (chair). Instead of directly showing the original ranking $\overrightarrow{p_{1-5\ on\ chair}}$, the system first takes into account the three sofas that have been chosen as user relevance feedback (see Section 5.1.6) and re-arranges the presentation order.

To apply the relevance feedback from the chosen sofas of customer C_{1-5} to generate a further personalized ranking order to present the chairs, we use association rules (see Section 5.2) to re-weight and rearrange the original $\overrightarrow{p_{1-5\ on\ chair}}$. In Equation 5.11, $\vec{p} = \overrightarrow{p_{1-5\ on\ chair}}$, and $\overrightarrow{p'} = \overrightarrow{p_{1-5\ on\ chair}}'$, where $\alpha = \beta = 1$. The new equation derived from Equation 5.11 is shown as below:

$$\overrightarrow{p_{1-5\ on\ cair}}' = \overrightarrow{p_{1-5\ on\ chair}} + \sum_{\forall \overrightarrow{d_j} \in D_r} \overrightarrow{d_j} \tag{6.1}$$

Meanwhile, Table 6-1 shows the list of index terms of the three sofas chosen by customer C_{1-5}. Parameter “frequency” means the number of times an index term appears in the whole user relevance feedback set (sofa14, sofa8, and sofa46). Therefore, since every sofa has 6 index terms (see Section 5.1.2) and 3 sofas are chosen, the total frequency is equal to 18 ($= 6 \times 3$).

After parsing the index terms from the three sofas, the system finds out the correlations of features between all previous choices (at this stage, sofas only) and the products in the next category (chairs). Then, Rule Group 1 is implemented before showing the chair page (see Section 5.2.2 and Figure 5-4). To search for applicable

Table 6-1: List of index terms of sofa8, sofa14, and sofa46 as relevance feedback

Index Terms	frequency
sofa_type_2	1
sofa_price_1	1
sofa_construction_1	1
sofa_comfort_3	2
sofa_color_2	3
sofa_design_4	1
sofa_type_3	1
sofa_price_3	1
sofa_construction_2	2
sofa_comfort_1	1
sofa_design_1	1
sofa_type_1	1
sofa_price_2	1
sofa_design_3	1

association rules from Rule Group 1, we consider the index terms from Table 6-1 as input X (see Table 5-6). Table 6-2 shows the results from the previous matching process. Parameter "frequency" indicates the importance of this rule and will be considered the weight.

Table 6-2: the association rules applicable for the three relevance feedback

X	Index Terms of X	Y	Index Terms of Y	support	frequency
Rule Group 1					
Sofa Design - Metal	sofa_design_4	Chair Design - Metal	chair_design_3	0.405797	1
Sofa Color - Dark Neutral	sofa_color_2	Chair Color - Dark Neutral	chair_color_1	0.434783	3
Sofa Comfort - Superior	sofa_comfort_3	Chair Comfort - Superior	chair_comfort_3	0.521739	2
Sofa Type - Casual	sofa_type_1	Chair Design - Medium Wood	chair_design_2	0.442029	1

After analysis, a set of relevant index terms of chairs (D_r) has been generated, as shown in Table 6-2. These terms are then re-weighted as shown in Equation 6.2.

$$\overrightarrow{p_{1-5\ on\ cair}}' = \overrightarrow{p_{1-5\ on\ chair}} + \sum_{\forall \overrightarrow{d_J} \in D_r} \overrightarrow{d_J} \tag{6.2}$$

where $\mathrm{D_r} = \{\text{chair\_design\_3}|\text{chair\_color\_1}|\text{chair\_comfort\_3}|\text{chair\_design\_2}\}$

6.2 Comparison

After obtaining the results of association rules, the new ranking ($\overrightarrow{p_{1-5\ on\ chair}}'$) for customer C_{1-5} can be calculated with Equation 6.2. The analysis shows that the three sofas selected by this customer C_{1-5} share the same index term "sofa_color_2", and that index term matches one association rule, which shows that in this demographic profile (Level 1-5), people who prefer sofas with attribute "sofa_color_2" as X are more likely to purchase chairs with attribute "chair_color_1" as Y (43% probability) (see Table 6-2). In order to observe the improvement, we highlight the chairs with attribute "chair_color_1" which is "Dark Neutral" in field Color.

The order of product presentation on the chair page for customer C_{1-5} is remarkably changed after personalization according to the preferences in the sofa category. To exemplify how the rankings are different, the original top 15 ranking orders $\overrightarrow{p_{1-5\ on\ chair}}$ (generated solely based on the demographic information of C_{1-5}) are shown in Table 6-3, with the new $\overrightarrow{p_{1-5\ on\ chair}}'$ shown in Table 6-4.

When this customer C_{1-5} selects one or multiple chairs and move on to the next group of product collection (table), this process will be applied the same manner. For example, $\vec{p}$ will be $\overrightarrow{p_{1-5\ on\ table}}$, $\overrightarrow{p'}$ will be $\overrightarrow{p_{1-5\ on\ table}}'$, and the feasible association rules will be from Rule Group 2 and Rule Group 3.

Table 6-3: The original $\overrightarrow{p_{1-5\ on\ chair}}$

Rank	ID	Name	Price	Construction	Comfort	Color	Design	Score
1	chair19	Casual_C1_M1	799	3	3	Light Neutral	Dark Wood	0.699846
2	chair24	Modern_C1_M3	799	2	3	Light Neutral	Metal	0.661431
3	chair23	Modern_C1_M2	599	2	3	Vivid Color	Metal	0.656987
4	chair32	Modern_C1_M2	399	2	2	Vivid Color	Metal	0.655437
5	chair10	Modern_C1_M3	599	3	2	Light Neutral	Metal	0.642899
6	chair22	Casual_C1_M3	399	2	3	Light Neutral	Medium Wood	0.634672
7	chair6	Casual_C1_M2	599	1	2	Light Color	Dark Wood	0.630066
8	chair30	Modern_C2_M3	999	2	3	Dark Neutral	Metal	0.62499
9	chair37	Modern_C2_M1	399	1	1	Vivid Color	Metal	0.615999
10	chair1	Modern_C2_M3	799	3	2	Dark Neutral	Metal	0.613251
11	chair20	Casual_C1_M2	999	3	1	Light Color	Dark Wood	0.609954
12	chair9	Casual_C1_M1	999	2	1	Light Neutral	Dark Wood	0.596406
13	chair11	Casual_C2_M2	999	1	3	Light Neutral	Medium Wood	0.593041
14	chair13	Trad_C1_M2	199	2	3	Dark Neutral	Medium Wood	0.59108
15	chair3	Casual_C2_M3	999	3	3	Medium Neutral	Medium Wood	0.582712

Table 6-4: The adapted $\overrightarrow{p_{1-5\ on\ chair}}'$

Rank	ID	Name	Price	Construction	Comfort	Color	Design	Score
1	chair19	Casual_C1_M1	799	3	3	Light Neutral	Dark Wood	0.526992
2	chair30	Modern_C2_M3	999	2	3	Dark Neutral	Metal	0.52439
3	chair23	Modern_C1_M2	599	2	3	Vivid Color	Metal	0.51593
4	chair13	Trad_C1_M2	199	2	3	Dark Neutral	Medium Wood	0.513958
5	chair5	Modern_C1_M3	999	2	1	Light Neutral	Metal	0.502099
6	chair6	Casual_C1_M2	599	1	2	Light Color	Dark Wood	0.496313
7	chair32	Modern_C1_M2	399	2	2	Vivid Color	Metal	0.492996
8	chair3	Casual_C2_M3	999	3	3	Medium Neutral	Medium Wood	0.488453
9	chair1	Modern_C2_M3	799	3	2	Dark Neutral	Metal	0.484601
10	chair33	Trad_C2_M3	999	3	3	Medium Color	Medium Wood	0.481729
11	chair20	Casual_C1_M2	999	3	1	Light Color	Dark Wood	0.47993
12	chair24	Modern_C1_M3	799	2	3	Light Neutral	Metal	0.479031
13	chair22	Casual_C1_M3	399	2	3	Light Neutral	Medium Wood	0.470293
14	chair37	Modern_C2_M1	399	1	1	Vivid Color	Metal	0.468112
15	chair11	Casual_C2_M2	999	1	3	Light Neutral	Medium Wood	0.4649

6.3 Discussion

It has been argued that the user profiles generated based only on demographics or stereotypes cannot provide adequate information about the user to make a precise and personalized prediction (Burke, 2002; Montaner, Lopez, & Rosa, 2003; Wei, Moreau, & Jenning, 2005). In the example discussed earlier in this chapter, despite the fact that users in "Profile 1-5" tend to select the chairs in a certain way, $\overrightarrow{p_{1-5\ on\ chair}}$, , this customer C_{1-5} might have different tastes or needs from the majority of other users in Profile 1-5. Therefore, with an attempt to overcome the generalization problem and cater to the user's individual needs, other techniques such as association rules from data mining have been employed in this system to find out the relationships between each choice and what product features are the primary factors in the user's purchase decision making. Using the same example from Section 6.2, because "chair_color_1" was found to be one of the critical product attributes, the system automatically reweights this index term and rearranges the presentation order of the chairs.

By comparing the original ranking and new ranking, listed in Table 6-3 and Table 6-4 respectively, we notice that chairs with color attributes "Dark Neutral" advance including "chair30" moves from top 8 to top 2; "chair1" from 9 to 8; "chair13" from14 to 4, which even outranks "chair1". This could be due to a match of another index term, e.g. "chair_comfort_3" as Y in finding the association rules (sofa_comfort_3 => chair_comfort_3).

The results demonstrate that while $\overrightarrow{p_{1-5\ on\ chair}}$ is made during the off-line phase, $\overrightarrow{p_{1-5\ on\ chair}}'$, the outcome of on-line analyses, can differ for each individual user and his/her dynamic interactivities. Our hybrid ranking system is self-learning and able to

personalize the product presentation gradually as the users interact and give feedback. As the number of users and interactions increases, the system will predict choices with more accuracy and will be more efficient with its recommendations.

Chapter 7

Conclusion and Future Work

7.1 Complete of work

This recommendation system is designed as an integration of the prevalent 3D virtual reality technology to establish a vivid and personalized shopping environment for customers. The recommendations are generated based on the users' demographic information as well as their dynamic navigation behaviors, such as product choices. Several techniques have been employed in this system including web usage mining, user profiling, hybrid ranking approaches, vector-spaced models and association rules.

Initially, we create multi-dimensional vector models to represent the associations between user profiles and products. Then, by performing the cosine measure on every pair of vectors to find their similarity, the system can successfully predict the user's interests and recommend products based on the quantitative scores. The higher the score of the products in our recommendation list, the more reliable it is to the users. In this project, our user profiling method maintains a hierarchical scheme to manage clustered groups based on their demographic information. Within these user profiles, we have successfully discovered some user shopping inclination from antecedent choices toward

further possible decisions by means of association rules.

This system also adjusts itself to satisfy users' individual needs by integrating their real-time responses as user relevance feedback with the specific recommendations generated from their profiles. On an e-commerce website, log data are often important indicators of the user's needs and interests. Therefore, by applying data mining techniques (such as association rules and sequential pattern discovery) to the log and user's personal data, some hidden but useful or interesting patterns can be found. We believe that this recommendation system, which is built upon both the user's personal data and dynamic behaviors, can adapt to each user progressively and finally achieve its maximum effectiveness.

7.2 Limitation

Despite the successful application on VRIS (http://www.vr-solution.com/), a furniture shopping website, the system has some inherited limitations. First of all, because some demographic data change from time to time (e.g., age and income), the information stored in user profiles may not be accurate until the user updates. In addition, 3D virtual reality technique requires more advanced graphical components of the computers. If a user visits this website on a computer that does not meet these requirements, the poor quality would largely reduce the precision of the user's decision or even prohibit the use of the system.

Running experiments with association rules and calculating the similarity of multi-dimensional vectors are time-consuming tasks that involve parsing and a prearranged data format. As the number of products increases, the number of attributes

(index terms), the number of correlation patterns, and the size of multi-dimensional vectors will have to be increased. Then, the latency problem (the system is not able to keep up with the pace of the user) may become an issue. In that case, the number of attributes extracted from products must be compressed into essentials only, rather than taking all the product features into consideration.

7.3 Future work

Currently, we create and manage our user profiles hierarchically based mainly on demographic information. However, our user matching techniques can be designed to expand the definition of personal information in user profiles by classifying and quantifying the conceptual aspects of users, such as lifestyles and personalities. Also, although the current system is a self-learning device, an optional profile training program, such as the one used in voice recognition technology, can be implemented with personal inquisition questionnaires so that the users could inform the system of his/her own preferences to mutually optimize the effectiveness of the recommendation system.

Additionally, in order to improve the quality of recommendations, a measurement to find out the variation between the ranking generated by the system and the user's real choices can be included to evaluate the performance and satisfaction. This could be done by measuring 1) the similarity between vectors of the adapted suggestions list and the user's product selection, 2) the ratio of final decision to the top 6 ranked items, and 3) the score of user final decision.

With the significant impact of 3D virtual reality on e-commerce, it is an ongoing project to explore every possibility to enhance the flexibility and convenience of

interactions with the system. This is because the more the shopping environment resembles the real world, the more precise the purchase decisions will be. Figure 7-1 presents the planning layout of our future website. By doing so, more useful and meaningful patterns and navigational trails could be revealed for further analysis.

Finally, in a study proposed by Granka, Joachims and Gay (2004), it is suggested that the relationship between how on-line users browse abstracts and how they select links for further exploration of an article can be exposed by tracking their eye movements. Similarly, this technique could be introduced into the system by tracking the consumers' eye movements to discover their favorite products or the features they interest them most. It is our ultimate goal to offer the customers a 3D virtual reality environment and dynamic adapting functions with the employment of multiple indicators of users' shopping behaviors to precisely personalize the e-commerce site for each individual.

Figure 7-1: the new interface of our website

REFERENCES

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Between Sets of Items in Massive Database. In *International Proceedings of the ACM-SIGMOD International Conference on Management of Data*, 207-216.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the International Conference on Very Large Data Bases*, 407-419.

Ansari, A., Essegaier, S. & Kohli, R. (2000). Internet Recommendation Systems. *Journal of Marketing Research, 37*(3), 363-375.

Ardissono, L. & Goy, A. (2000). Tailoring the Interaction with Users in Web Stores. *User Modeling and User-Adapted Interaction 10*(4), 251-303.

Ardissono, L., Goy, A., Meo, R., Petrone, G., Console, L., Lesmo, L., Simone, C., Torasso, P. (1999). A Configurable System for the Construction of Adaptive Virtual Stores. *World Wide Web, 2*(3), 143-159.

Chen, L., Gillenson, M., & Sherrell, D. (2004). Consumer Acceptance of Virtual Stores: A Theoretical Model and Critical Success Factors for Virtual Stores. *ACM SIGMIS Database, 35*(2), 8-31.

Fiore, A., and Jin, H. (2003). Influence of Image Interactivity on Approach Responses Towards an Online Retailer. *Internet Research, 13*(1), 38-48.

Edwards, S. & Harshavardhan, G. (2001). The Novelty of 3D Product Presentations Online. *Journal of Interactive Advertising, 2*(1). 22-32.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Boston, MA: Addison-Wesley Longman Publishing Co., Inc..

Balabanovic, M. & Shoham, Y. (1997). Content-based Collaborative Recommendation. *Communications of the ACM, 40*(3), 66-72.

Burke, R. (2002). Hybrid Recommender System: Survey and Experiments. *User Modeling and User-Adapted Interaction, 12*(4), 337-370.

Chitttaro, L. & Ranon, R. (2002). Dynamic Generation of Personalized VRML Content: a General Approach and its Application to 3D E-Commerce. In *Proceedings of 7th International Conference on 3D Web Technology, Web3D'2002*, ACM Press, 145-154.

Cole, J. I., Suman, M., Schramm, P., Lunn, R., & Aquino, J. S. (2003). Surveying the Digital Future-year Three. *The UCLA Internet Report, UCLA Center for Communication Policy*, 42-46.

Cristofor, L. (2006). "ARMiner Info Index". Retrieved July 15, 2007, from World Wide Web: http://www.cs.umb.edu/~laur/ARMiner/.

Dunham, M. (2003). *Data Mining: Introductory and Advanced Topics*. Upper Saddle River, NJ: Prentice Hall.

Edwards, S. & Gangadharbatla, H. (2001). The Novelty of 3D Product Presentations Online. *Journal of Interactive Advertising, 2*(1), 22-32.

Eirinaki, M & Vazargiannis, M. (2003). Web Mining for Web Personalization. *ACM Transactions on Internet Technology, 3*(1), 1-27.

Granka, L., Joachims, J., & Gay, G. (2004). Eye-tracking Analysis of User Behavior in WWW Search. In *Proceedings of SIGIR'04*. ACM Press. 478-479.

Haubl, G. & Trifts, V. (2000). Consumer Decisoin Making in Online Shopping Environments: The Effects of Interactive Decision Aids. *Marketing Science, 19*(1), 4-21.

Kobsa, A., Koenemann, J., & Pohl, W. (2001). Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships. *The Knowledge Engineering Review, 16*(2), 111-155.

Krulwich, (1997). Lifestyle Finder: Intelligent User Profiling Using Large-scale Demographic Data. *Artificial Intelligence Magazine, 18*(2), 37-45.

Lee, W., Liu, C., Lu, C. (2002), Intelligent Agent-based Systems for Personalized Recommendations in Internet Commerce. *Expert Systems with Applications, 22*, 275-284.

Lee, D., Chuang, H., & Seamons, K. (1997). Document Ranking and the Vector-Space Model. *IEEE Software, 14*(2), 67-75.

Lekakos, G. & Giaglis, G. *(2006). Improving the Prediction Accuracy of Recommendation Algorithms: Approaches Anchored on Human Factors. Interacting with Computers, 18*(3), 410-431.

Lekakos, G. & Giaglis, G. (2007). A Hybrid Approach for Improving Predictive Accuracy of Collaborative Filtering Algorithms. *User* Modeling and User-*Adapted Interaction, 17*(1-2), 5-40.

Li, Hairong, Daugherty, Terry, & Biocca, Frank (2002). Impact of 3-D Advertising on Product Knowledge, Brand Attitude, and Purchase Intention: The Mediating Rile of Presence. *Journal of Advertising, 31*(3), p43-57.

Li, H., Daugherty, T., & Biocca, F. (2001). Characteristics of virtual experience in electronic commerce: A protocol analysis. *Journal of Interactive Marketing, 15*(3), 13-30.

Montaner, M., Lopez, B., & Rosa, J. (2003). A Taxonomy of Recommender Agents on the Internet. *Artificial Intelligence Review, 19*(4), 285-330

Pazzani, M. (1999). A Framework for Collaborative, Content-Based and Demographic Filtering. *Artificial Intelligence Review, 13*(5/6), 393-408.

Pazzani, M. & Billsus, D. (1997). Learning and Revising User Profiles: The Identification of Interesting Web Sites. *Machine Learning, 27*, 313-331.

Pennington, R. (2001). Signs of Marketing in Virtual Reality. *Journal of Interactive Advertising, 2*(1), 43-55.

Putrevu, S. & Lord, K. R. (2003). Processing Internet Communications: A Motivation, Opportunity and Ability Framework. *Journal of Current Issues and Research in Advertising, 25*(1), 45-59.

Salton, G. (1971). *The SMART Retrieval System – Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice Hall Inc..

Salton, G. & Buckley, C. (1988). Term-weighting Approaches in Automatic Retrieval. *Information Processing & Management, 24*(5), 513-523.

Schafer, J. B., Konstan, J. A., & Riedl, J. (1999). Recommender Systems in E-Commerce. In *EC '99: Proceedings of the First ACM Conference on Electronic Commerce, Denver, CO.*, 158-166.

Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-Commerce Recommendation Applications. *Data Mining and Knowledge Discovery, 5*(1/2), 115-153.

Senecal, S. & Nantel, J. (2004). The Influence of Online Product Recommendations on Consumers' Online Choices. *Journal of Retailing, 80*, 159–169.

Smith, D., Menon, S., & Sivakumar, K. (2005). Online Peer and Editorial Recommendations, Trust, and Choice in Virtual Markets. *Journal of Interactive Marketing, 19*(3), 15-37.

Thearling, K. (1995). From Data Mining to Database Marketing. Retrieved July 15, 2007, from World Wide Web: http://www.thearling.com.

Vozalis, M. & Margaritis, K. (2004). Collaborative Filtering Enhanced by Demographic Correlation. In *Proceedings of the AIAI Symposium on Professional Practice in AI, Part of the 18th World Computer Congress, Toulouse, France*, 393–402.

Wei, Y. Z., Moreau, L., & Jennings, N. R. (2005). A Market-Based Approach to Recommender Systems. *ACM Transactions on Information Systems, 23*(3), 227-266.