Abstract

The Internet revolution changed the world and made it as a small village, since everyone can contact people anywhere in the world. This easy communication facilitates selling and buying through the Internet which is called e-commerce.

When e-commerce began to grow, problems appeared, one of them is how to buy something from a huge category, i.e. when a customer wants to buy something from internet markets, he will be confused what to choose and from where, because of the various items and enormous sites.

People handle this information overload through their own effort, the effort of others and some blind luck. First of all, most items and information are removed from the stream simply because they are either inaccessible or invisible to the user. Second, a large amount of filtering is done for us. Newspaper editors select what articles their readers want to read. Bookstores decide what books to carry. However with the dawn of the electronic information age, this barrier will become less and less a factor. Finally, we rely on friends and other people whose judgement we trust to make recommendations to us.

A technology is needed to help people wade through all the information to find the items they really want and need, and to rid them of the things they do not want to be bothered with.

Recommender systems are the new technology that assist and augment the recommendation process. In a typical recommender system people provide recommendations as inputs, which the system then aggregates and directs to appropriate recipients.

In this work, a recommender system is built that uses different recommendation methods.

الخلاصة

لقد غيرت ثورة الانترنيت العالم بأسره فأصبح قرية صغيرة و صار بإمكان أي شخص في أي مكان أن يطوف العالم بأسره و هو جالس أمام شاشة الكومبيوتر. من هنا ظهر الشراء و البيع عن طريق الانترنيت أو ما يعرف الآن بالتجارة الإلكترونية (E-commerce).

هذا التوسع خلق مشكلة جديدة و هي صعوبة الاختيار. فصار الزبون يشعر بالحيرة حينما يريد أن يختار حاجةً ما إذ أنه أمام آلاف المواقع التي توفر الحاجة التي يريدها و كل موقع منها يعرض أشكالاً و أنواعا مختلفةً للحاجة الواحدة. من هنا ظهرت الحاجة لتقنية جديدة تتمثل اليوم في الأنظمة الناصحة.

إننا في كثير من الأحيان نقوم بعمل ما دون أن نملك الخبرة الكافية للقيام به. و في حياتنا اليومية كلنا يعتمد من وقت لآخر على نصائح مختلفة إما من ناس يخبروننا بها أو عن طريق وسائل الإعلام المختلفة أو من مسؤولين عن الخدمات العامة.

النظام الناصح يساعد و ينمي هذه العملية الطبيعية التي نقوم بها يومياً في مجتمعاتنا. في النظام الناصح يدخل المستخدمون نصائحهم فيقوم النظام بجمعها و توجيهها إلى المتلقي.

يعمل النظام الناصح بأن يسألك سلسلة من الأسئلة عن الأشياء التي تحبها و التي لا تحبها. ثم يقارن إجاباتك بإجابات زبائن غيرك و من خلال هذه المقارنة يجد أقرب الزبائن إليك من خلال الشبه في الآراء. و بهذا ينصحك بما يحب الزبون القريب منك.

في هذه الأطروحة تم بناء نظام ناصح يعتمد على مجموعة من الطرق لكي يقدم النصائح للمستخدمين.

Chapter One

Introduction

1.1 Introduction

The first recorded description of the social interactions that could be enabled through networking was as memos written in August 1962 discussing "Galactic Network" concept describing a globally interconnected set of computers through which everyone could quickly access data and programs from any site. In spirit, the concept was very much like the internet of today. By 1985, Internet was already established as a technology supporting a broad community of researchers and developers, and was beginning to be used by other communities, often with different systems. The first commercial announcement was at 1988, which considered to be the start of what will be called "e-commerce". As the new revolution -the Internet- developed, the whole world changed and became as a small village, since everyone can contact people everywhere in the world. That easy communication made e-commerce spread and broadly used. Nowadays, it is not an easy mission to choose an item from Internet. Data Mining is introduced as a solution. [28]

Data Mining is a field of knowledge discovery used to predict with some accuracy when generators are likely to fail. The technique started making more inroads into the corporate world in the 1990s, catching on as a means to detect fraud in the insurance, health care and credit card industries. By finding patterns and predicting likely behavior, companies can catch people who lie on applications or are likely to engage in dangerous or illegal activities.

Department stores, supermarkets and other brick-and-mortar retailers have used data mining to guess customer buying habits for years, but relatively few general consumer e-tailers and content producers have fully

exploited the research technique. That's partly because the practice--involving algorithms, samplings and parallelisms--is complicated and poorly understood. But it's starting to find its way into the mainstream.

Data mining introduce algorithms to guess customer opinions and then recommend the customer with items requested. This lead to recommender systems. [14]

The necessity of system-originated recommendation becomes imperative as human computer interaction becomes more complicated. Software packages are becoming more complex, the number of provided services is increasing, the range of selection is widening, and the user, in general is confronted with numerous dilemmas. Recommender Systems can guide the user through these processes by recommending paths, solutions, alternatives and new ideas. Recommendation in most cases is regarded as an extension of the prediction process that frequently takes place in user modeling systems– instead of predicting a single item of a user's profile based on other information, a whole set of items is predicted in a similar fashion. The predicted items will thus play the role of items to be recommended to the user. [23]

1.2 E-Commerce

Electronic commerce (or e-commerce) is defined as the conduct of commerce in goods and services, with the assistance of telecommunications and telecommunications-based tools. [29]

E-Commerce is about setting your business on the Internet, allowing visitors to access your web site, and go through a virtual catalog of your products/ services online. When a visitor wants to buy something he/she likes, they merely, "add" it to their virtual shopping basket. Items in the virtual

shopping basket can be added or deleted, and when the customer is all set to checkout...he head to the virtual checkout counter, which has his complete total, and will ask him some personal information such as name, address etc. and the method of payment (usually via credit card). Once the customer had entered all this information (which by the way is being transmitted securely) he can then just wait for delivery. It's that simple.

E-Commerce is not about just online stores, it's about anything and everything to do with money. If the customer pay (via cash, check, credit card, etc.) Days are not far away when anyone would be able to order and reserve a request for any item at a store (all online). [31]

"E-commerce is the newest and hottest use," said Michael Gilman, president and chief executive of Data Mining Technologies. "Anywhere you have historical data, you can use it to get patterns that you can't see with the human eye." [25]

1.2.1 E-Commerce Advantages

First, on the web, data are collected electronically rather than manually so less noise is introduced from manual processing, which will increase data security and validity. To understand this point one can compare between e-commerce and the traditional commerce where people buy from real stores and pay physical money. In traditional commerce, sellers or buyers might cheat, in e-commerce, cheating is less, because as mentioned before, in e-commerce the whole operation will be done electronically so it will be under the control of the system. Second, electronic data are rich, containing information on prior purchase activity and detailed demographic data. In addition, some data that previously were very difficult to collect now are accessible easily. For example, electronic commerce systems can record the actions of customers in the virtual store. Also for electronic commerce

systems massive amount of data can be collected inexpensively. The electronic commerce system is easy to implement and evaluate data mining models because the Internet already is automated. [9]

Finally, implementing E-Commerce saves money, and customers can buy from electronic stores online 24 hours-a-day, 7 days a week, with no traffic jams, shopping crowds, carrying overloaded heavy shopping bags etc. [31]

The e-commerce process activities include:

1. Electronic presentation of goods and services
2. Online order taking and bill presentment
3. Automated customer account inquiries
4. Online payment and transaction handling [32]

1.2.2 Core Components Of any E-commerce Web Site

Any E-Commerce web site is basically involve combining an easy-to-use, manageable web site design with a Shopping Cart Program and an Online Merchant Account -- then setting those up through a reliable E-Commerce Hosting provider. Figure 1.1 illustrates the relationships between e-commerce components. [1]

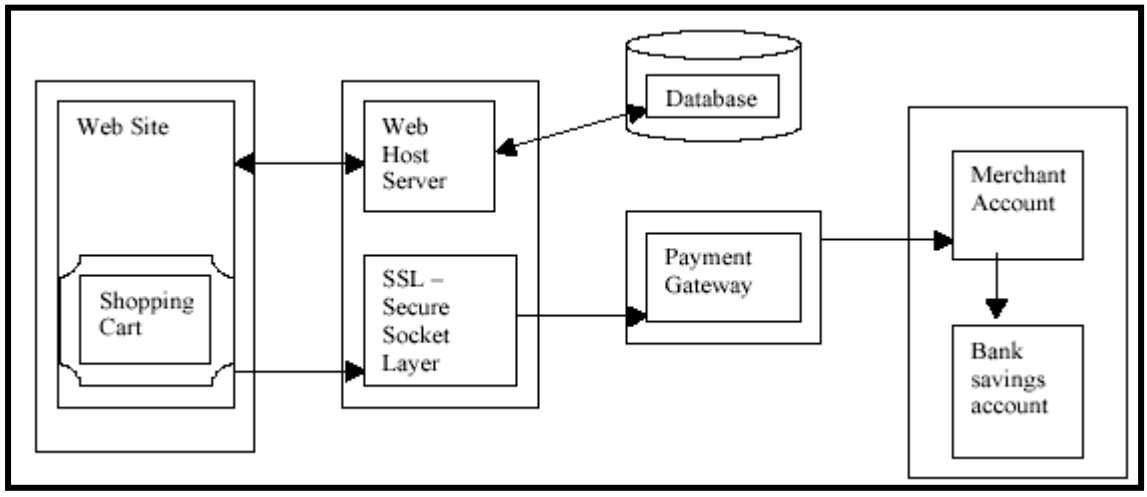

Figure 1.1: Basic diagram of an e-commerce web site

Web Site [8]

The first component the designer obviously need is a Web site. Web site will have dynamic web site capability (involving a database). This is the basic foundation of a web site. It is like building a house, a web site have a foundation, walls and a roof. Any upgrades can only be added after this point. This is where the marketing aspects can be applied. How to get customers to buy? What should the content be for the web site?.. etc.

Shopping Carts [8]

The next step to get the web site to sell is to add a shopping cart. A shopping cart is a web application that allows multiple simultaneous users to select specific items for purchase, while keeping a running total of combined item and processing costs. The end result is the total costs of items and processing charges are charged to the customer, and all customer and order information is available to seller, so the purchase can be completed.

Database [8]

The database holds the numerous data information that is collected during the shopping process and it can also hold other types of information that helps make the site function.

Secure Socket Layer (SSL) [8]

If the designer choose to accept credit cards on the site, there will be a need to understand the Secure Socket Layer (SSL).

SSL is a public/private key encryption process. This allows the sensitive data the user puts into the form to be sent to the server in an encrypted format. It basically secures the data from the shopper's browser to the web server.

Payment Gateway [8]

This is a service/program that collects the sales amount, and the credit card information from the customer and sends the payment to the bank via the Internet.

The payment gateway needs to be connected to the web site and its bank account over the Internet. It acts as a money conduit between the shopper and the business bank account.

Merchant Account [8]

A merchant account allows the site business to charge sales to credit cards. It's an account/agreement between the site business, and a bank (in cooperation with Visa, MasterCard, American Express, etc.). Any business that accepts credit cards (on or off line) must have a Merchant Account

Business Bank Account [8]

This isn't just required for businesses accepting credit cards, but for any business.

However, the designer will need a bank account that will accept the deposits for the credit card charges.

Putting them all together into a working E-Commerce web site does require some technical knowledge. [1]

1.2.3 Types of E-commerce [12]

There are four generally accepted types of e-commerce:

1. Business to Business (B2B): In this type, e-commerce is done between business non-governmental institutions. Example, a company buys from an e-commerce web site for a factory.
2. Business to Consumer (B2C): This is the most common type for Internet users. In which individual consumers buy from business e-commerce web site. Example, someone buys an item from a business web site.

3. Government to Business: This type the government deals with non-governmental institutions. Example, governmental tenders is a common example where government institutions make tenders for particular items and companies or factors provide these items.
4. Government to Citizen: This type is similar to B2C type except that the institution is governmental. Example, a governmental factory produces items for consumers.

1.2.4 E-commerce Servers [21]

Electronic commerce server can be best defined as a Web software that runs some of the main functions of an on-line storefront such as product display, on-line ordering, and inventory management.

The first commerce servers were developed by IBM, Netscape and Open Market. Since then companies as iCat, Inex, Microsoft, Connect, Oracle and Viaweb have developed commerce servers also.

1.2.5 E-commerce Payment Systems [2]

A payment system simply transfers digital representations of funds from one computer to another. Like serial numbers on real dollar bills, the digital cash numbers are unique identifier carrying a given value, while each one is issued by a participating bank and represents a special sum of real money. One of its key features is that the real cash is anonymous and reusable. (As opposed to the credit cards)

1.3 Recommender Systems [7]

Electronic commerce systems allow unprecedented flexibility in merchandising. However, flexibility is not a benefit unless one knows how to map the many options to different situations. Because of this flexibility, the one can not decide easily what to buy from a huge amount of products. For

example, how should a buyer choose a particular shirt from tens of sites that contains hundreds of shirts with different colors, prices, sizes, and models?

Recommender systems apply data mining techniques to the problem of making personalized recommendations for information, products or services during a live interaction. These systems are achieving widespread success on the Web. The tremendous growth in the amount of available information and the number of visitors to Web sites in recent years poses some key challenges for recommender systems. These are: producing high quality recommendations, performing many recommendations per second for millions of users and items and achieving high coverage in the face of data scarcity. In traditional collaborative filtering systems the amount of work increases with the number of participants in the system. New recommender system technologies are needed that can quickly produce high quality recommendations, even for very large-scale problems.

Recommender systems apply data analysis techniques to the problem of helping users find the items they would like to purchase at E-Commerce sites by producing a predicted likeliness score or a list of top–N recommended items for a given user. Item recommendations can be made using different methods. Recommendations can be based on demographics of the users, overall top selling items, or past buying habit of users as a predictor of future items. Collaborative Filtering (CF) is the most successful recommendation technique to date. The basic idea of CF-based algorithms is to provide item recommendations or predictions based on the opinions of other like-minded users. The opinions of users can be obtained explicitly from the users or by using some implicit measures.

1.4 Aim Of This Thesis

The main goal of this thesis is to build a recommender system to advice a customer the best items that suit his/her interest from a selected site. Different algorithmic strategies are going to be used and a comparison study would be made between these different recommendation algorithms.

An existing movie society data base consists of different existing and proposed algorithms.

And finally, this thesis leads the way to other future recommender systems in Iraq by telling the best algorithms and ways to recommend.

1.5 Related Work [15]

Six E-commerce businesses that use one or more variations of recommender system technology in their web sites will be presented. For each site, and each variation, a brief description of the features of the system is given. For organizational purposes these sites have been alphabetized. The descriptions of these sites are accurate up to the time of writing this thesis, though E-commerce applications of recommender systems are changing rapidly.

a. Amazon.com

First, a user have to sign up first as a new customer (if he was not already a customer) by filling personal information. Each customer will have a virtual cart that contains items the customer selected to buy.

Like many E-commerce sites, Amazon.com™ (www.amazon.com) is structured with an information page for each item, giving details of the text and purchase information. First, Amazon shows a list of top-selling items, either they are frequently purchased by customers or highly recommended

from other customers. Amazon also encourages direct feedback from customers about items they have purchased. Customers rate books they have read on a 5-point scale from “hated it” to “loved it.” After rating a sample of items, customers may request recommendations for items that they might like. At that point, a half dozen non-rated texts are presented that correlate with the user’s indicated tastes. Amazon also has the Eyes feature which allows customers to be notified via email of new items that have been added to the Amazon.com catalog. And finally, the Customer Comments feature allows customers to receive text recommendations based on the opinions of other customers. Located on the information page for each book is a list of 1-5 star ratings and written comments provided by customers who have read the book in question and submitted a review. Customers have the option of incorporating these recommendations into their purchase decision. Furthermore, customers can “rate the comments.” With each comment is the question “Did this comment help you.” Customers may indicate yes or no. Results are tabulated and reported such as “5 of 7 people found the following review helpful.”

b. CDNOW

Customers locate the information page for a given album or artist. The system then recommends ten other albums related to the album or artist in question. Results are presented as “Customers who bought X also bought set S” or “Customers who bought items by Y also bought set T.” Customers type in the names of up to three artists, and the system returns a list of ten albums CDNOW considers similar to the artists in question. The Related Artists feature of CDNOW works on the assumption that if a customer likes a certain performer, there is a group of artists with similar styles that he will also like. The Buyer’s Guide feature at CDNOW allows customers to receive

recommendations based on a particular genre of music. Customers browse a list of genres provided by the site, including categories. Selecting one of the links from this list takes customers to a new list of albums the editors consider the essential part of this genre.

Top 100: Traditionally, “bestseller” status have been used by commerce sites to make recommendations to their customers.

c. My CDNOW

My CDNOW enables customers to set up their own music store, based on albums and artists they like. Customers indicate which albums they own, and which artists are their favorites. Purchases from CDNOW are entered automatically into the “own it” list. When customers request recommendations, the system predicts six albums the customer might like based on what is already owned. Feedback is provided by customers. The albums recommended change based on the feedback.

d. Drugstore.com

The Advisor feature at Drugstore.com allows customers to indicate their preferences when purchasing a product from a category such as “suncare” or “cold and flu remedies.” For example, in the latter, customers indicate the symptoms they wish to relieve (runny nose and sneezing), the form in which they want the relief (caplets) and the "age” of patient to whom they want to administer the product (adult). Upon being provided with this information the Advisor returns a list of products recommended to meet the conditions. In the Test Drives feature, a team of volunteers, made up of customers from the site, is sent a new product. These “fellow customers” provide reviews of the product including a star rating and text comments.

e. eBay

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

Customers input a "short term" (30/60/90 days) and search on a set of keywords of their choosing, including their price limit. On a periodic basis (one or three day intervals) the site performs the customer's search over all auctions at the site and sends the customer an email with the results of this search.

f. MovieFinder.com

MovieFinder.com is the movie site maintained by E! Online. Both the Users Grade and the Our Grade features report a letter grade recommendation to the customer. The Users Grade feature allows customers to register with the site and give letter grades (A-F) to the movies they have seen. These grades are then averaged over all customers and reported as the Users Grade. The Our Grade feature provides customers with a grade from the editors of E! Online. Thus, customers viewing the information page for Toy Story 2 might find that it gets a grade of A from the editors with a grade of A- from the customers who have rated it. The Top 10 feature at E! Online allows the customers to get recommendations from the editors in a category of their

choice. Customers select a category from a list of previously defined categories such as chick flicks, and movies from books.

Selecting a list takes the customer through descriptions of the top ten movies in that category as defined by one of the editors of E! Online.

g. Reel.com

Reel.com's Movie Matches feature (www.reel.com) provides recommendations on the information page for each movie. These recommendations consist of "close matches" and/or "creative matches." Each set contains up to a dozen hyperlinks to the information pages for each of these "matched" films. The hyperlinks are annotated with one-sentence descriptions of how the new movie is similar to the original movie in question.

1.6 Thesis Layout

Chapter two explains recommender systems: what they mean, why they are used, and describes different recommendation algorithms and techniques.

Chapter three presents the recommendation methods and the proposed recommender system, also, explains the experimental database and displays the results gained from the proposed recommender system and discusses them.

Chapter four contains conclusions and suggestions for further work.

Chapter Two

The Recommender Systems in E-Commerce

2.1 Introduction

The problem of predicting a user's behavior on a web-site has gained importance due to the rapid growth of the world-wide-web and the need to personalize and influence a user's browsing experience. [20]

It is often necessary to make choices without sufficient personal experience of the alternatives. In everyday life, we rely on recommendations from other people either by word of mouth, recommendation letters, movie and book reviews printed in newspapers, or general surveys.

Recommender systems assist and augment this natural social process. In a typical recommender system people provide recommendations as inputs, which the system then aggregates and directs to appropriate recipients. In some cases the primary transformation is in the aggregation; in others the system's value lies in its ability to make good matches between the recommenders and those seeking recommendations. The developers of the first recommender system, Tapestry, coined the phrase "collaborative filtering" and several others have adopted it. The more general term "recommender system" is preferred for two reasons. First, recommenders may not explicitly collaborate with recipients, who may be unknown to each other. Second, recommendations may suggest particularly interesting items, in addition to indicating those that should be filtered out.[24]

A recommender system works by asking a series of questions about things the customer liked or didn't like. It compares customer's answers to others, and finally recommends depending on people opinions. [16]

2.2 Data Mining

Data Mining is the process of extracting of, previously unknown, valid, and actionable information from large database(s) and then using the information to make important business decisions. [10]

The underlined words in the definition lend insight into the essential nature of data mining and help to explain the fundamental differences between it and the traditional approaches to data analysis such as query and reporting and Online Analytical Processing (OLAP). In essence, data mining is distinguished by the fact that it is aimed at the discovery of information, without a previously formulated hypothesis.

First, the information discovered must have been previously unknown. Although this sounds obvious, the real issue here is that it must be unlikely that the information could have been hypothesized in advance;

That is the data miner is looking for something that is not intuitive or, perhaps, even counterintuitive. The further away the information is from being obvious potentially the more value it has. Data mining can uncover information that could not even have been hypothesized with earlier approaches.

Second, the new information must be valid. This element of the definition relates to the problem of overoptimism in data mining; that is, if data miners look hard enough in a large collection of data, they are bound to find something of interest sooner or later. For example, the potential number of associations between items in customers' shopping baskets rises exponentially with the number of items. Some chains carry upwards of 300,000 items, so the chances of getting spurious associations is quite high.

The possibility of spurious results applies to all data mining and highlights the constant need for post-mining validation and sanity checking.

Third, and most critically, the new information must be actionable, that is, it must be possible to translate it into some business advantage. In many cases, however, the actionable criterion is not so simple. For example, mining of historical data may indicate a potential opportunity that a competitor has already seized. Equally, exploiting the apparent opportunity may require use of data that is not available or not legally usable. Needles to say, an organization must have the necessary political will to carry out the action implied by the mining.

The ability to use the mined data to inform crucial business decisions is another critical environmental condition for successful commercial data mining, and underpins data mining's strong association with any applicability to business problems. Figure 2.1 Shows a general positioning of the components in a data mining environment. [10]

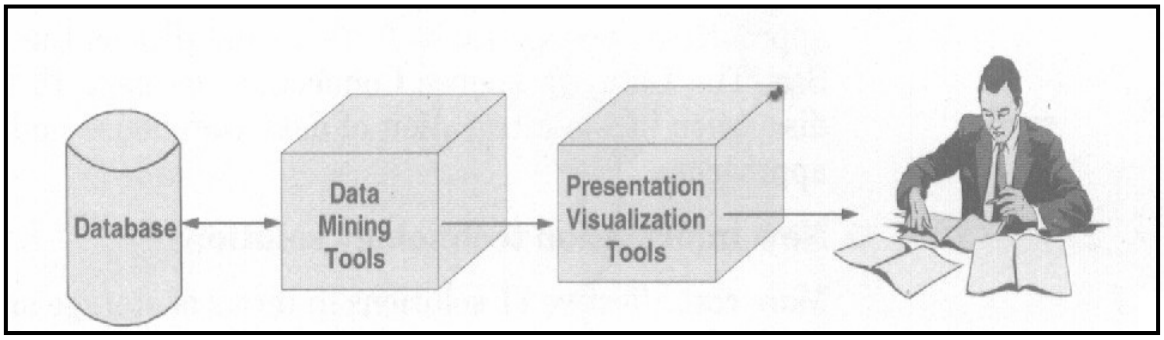

Figure 2.1: Data Mining Positioning

Data mining is often one or more of the following:

1- Structured query language (SQL) queries against a large database.

2- Advanced information retrieval, for example, through intelligent agents.

3- Multidimensional database analysis (MDA).

4- Online Analytical Processing (OLAP).

5- Exploratory data analysis (EDA).

6- Advanced graphical visualization.

7- Traditional statistical processing against a data warehouse.

None of these approaches is data mining because each lacks the essential ingredient- discovery of information without a previously formulated hypothesis. Figure 2.2 shows these approaches that are not data mining.

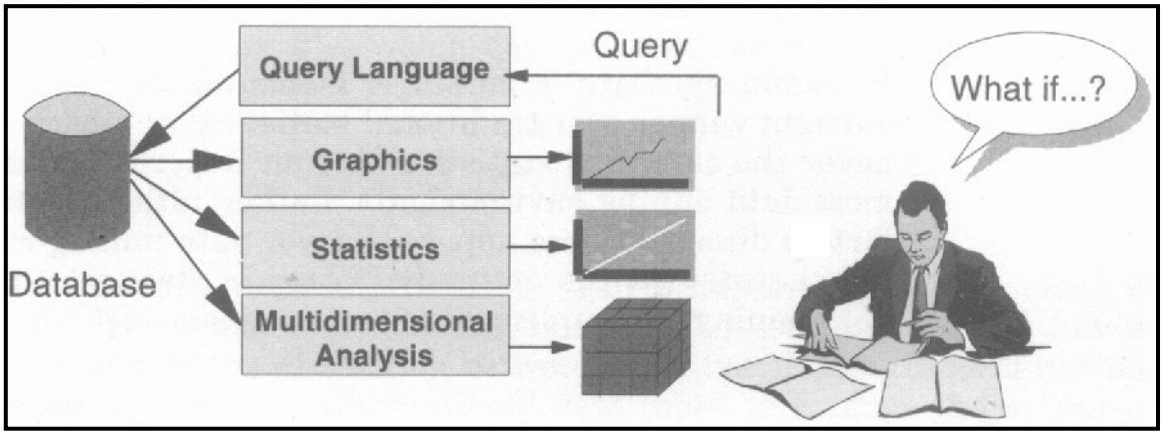

Figure 2.2: Traditional Data Analysis, Not Data Mining

2.2.1 Reasons for the growing popularity of Data Mining [19]

Growing Data Volume

The main reason for necessity of automated computer systems for intelligent data analysis is the enormous volume of existing and newly appearing data that require processing. The amount of data accumulated each day by various business, scientific, and governmental organizations around the world is daunting. Only scientific organizations store each day about 1 TB (terabyte!) of new information. And it is well known that academic world is by far not the leading supplier of new data. It becomes impossible for human analysts to cope with such overwhelming amounts of data.

Limitations of Human Analysis

Two other problems that surface when human analysts process data are the inadequacy of the human brain when searching for complex multifactor dependencies in data, and the lack of objectiveness in such an analysis. A human expert is always a hostage of the previous experience of investigating other systems. Sometimes this helps, sometimes this hurts, but it is almost impossible to get rid of this fact.

Low Cost of Machine Learning

One additional benefit of using automated data mining systems is that this process has a much lower cost than hiring an army of highly trained (and paid) professional statisticians. While data mining does not eliminate human participation in solving the task completely, it significantly simplifies the job and allows an analyst who is not a professional in statistics and programming to manage the process of extracting knowledge from data.

2.3 E-Commerce

E-Commerce can be defined as business activities conducted using electronic data transmission via the Internet. [17]

E-commerce is growing fast, and with this growth companies are willing to spend more on improving the online experience. Data Mining tools aid the discovery of patterns in customer data. [18]

Data mining techniques have much wider applicability than searching for patterns within customer data. They are also essential tools in the task of searching and extracting information from the Web. [27]

One of the applications resulted from using data mining algorithms in e-commerce web sites is recommender systems.

2.4 Recommender Systems

Recommender Systems advise their users about which items (products, services or information) to consume. [27]

With the rapid growth in size and number of available databases used in commercial, industrial, administrative and other applications, it is necessary and interesting to examine how to extract knowledge automatically from huge amount of data. Traditionally database systems were used to support business data processing applications, and much Database Management Systems research was focused in this direction. However, an important new use for the technology have recently emerged-which is known as “Data Mining”. There is today a great deal of enthusiasm for data mining, the extraction of (hidden) information from large bodies of data often accumulated for other purposes. Data Mining (sometimes called “ knowledge discovery in database”) has been considered as one of the most important research topics in database by many database researchers. [30]

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if it not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.[22]

Recommender systems are new types of internet-based software tools, designed to help users find their way through today’s complex on-line shops and entertainment web sites. [26]

The rapid expansion of the Internet has brought about a new market for trading. Electronic commerce or e-commerce has enabled businesses to open up their products and services to a massive client base that was once available only to the largest multinational companies. As the competition between businesses becomes increasingly fierce, consumers are faced with a myriad of choices. Although this might seem to be nothing but beneficial to the consumer, the sheer wealth of information relating to the various choices can

be overwhelming. One would normally rely on the opinions and advice of friends or family members but unfortunately even they have limited knowledge. [26]

Recommender systems provide one way of circumventing this problem. As the name suggests, their task is to recommend or suggest items or products to the customer based on his/her preferences. These systems are often used by E-commerce websites as marketing tools to increase revenue by presenting products that the customer is likely to buy. An internet site using a recommender system can exploit knowledge of customers' likes and dislikes to build an understanding of their individual needs and thereby increase customer loyalty. [26]

2.4.1 Recommendation Operation: [3]

1- Selection: The data set used to produce recommendations can stem from various sources. For example these sources can be already existing transaction logs (e.g. point of sale data, Web server logs) or the data can be collected specially for the purpose of generating recommendations (e.g. ratings).

2- Preprocessing and transformation: In these steps the data set is cleaned from noise, inconsistent data is removed and missing data is inferred. After this treatment the cleaned data is transformed into a representation suitable for data mining. For example, for collaborative filtering the data normally consists of explicit ratings by users that are collected for the purpose of creating recommendations. Preparation mainly involves discovering and removing inconsistent ratings. For Web usage mining, data is collected by observing the behavior of users browsing a Web site. Since observation, especially server-side observation on the Web, is far from perfect, much

effort has to be put on data preprocessing, cleaning and transformation. Problems involve the identification of users, missing data due to proxy servers and system crashes, requests by Web robots and many more.

3- Data mining: The objective of this step is to find interesting patterns in the data set that are useful for recommendation purposes. The output of data mining in recommender systems can be: groups of users with similar interests, items that are frequently used together, often used sequences of items,... Frequently, extracting patterns means learning the parameters of a specified model from the data set.

4- Interpretation and evaluation: In order to build knowledge, the found patterns (the model and its parameters) have to be understandable to humans. Only with this property the process can be called knowledge discovery and the results can be interpreted. A recommender system interprets found patterns for the user. Finally the validity (patterns are also valid for new data), novelty (involves a new way of finding patterns), usefulness (potentially lead to useful action) and understandability (build and increase knowledge) of the patterns needs to be evaluated.

5- Presentation: A recommender system presents this interpretation in a suitable form as a recommendation. For example, the recommendation can be a top-n list of recommended items for a user, or a list of items that are similar to an item the user likes, or it can consist of information about how other users with similar interests rated a specific item.

Figure 2.3 illustrates the recommendation operations and the relationships between them.

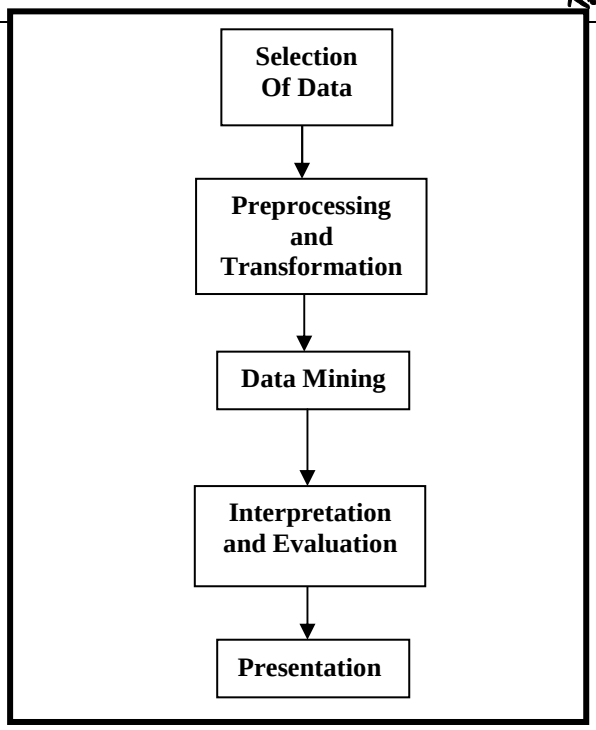

Figure 2.3 Recommendation Operation

2.4.1.1 Rating [13]

By rating is meant services by which the selection of resources to read is guided by the quality of the resources, as specified by people who have read the resource. Rating is also known under the terms "collaborative filtering" or "social filtering". In the Internet, rating may be applied to many kinds of resources, like web pages, messages, electronic journal papers, public domain software. The purpose of rating may be to increase the quality of the resources read, or to avoid certain resources deemed unsuitable in certain communities for certain groups of readers (example: violence, pornography).

In the world before the Internet, rating was commonly provided by services such as:

1. Newspapers, magazines, books, which are rated by their editors or publishers, selecting information which they think their readers will want.

2. Consumer organizations and trade magazines which evaluate and rate products.
3. Published reviews of books, music, theatre, films, etc.
4. Peer review method of selecting submissions to scientific journals.

2.4.1.2 Rating Disadvantages [13]

Some problems which can cause rating to work less well are:

1. Too few ratings are provided to provide a good basis for rating.

2. It may be difficult to collect ratings from users. Some systems solve this by implicitly guessing user ratings from the time the user spends reading a resource.

3. Some raters may not do a good work of rating. This could happen when people don't care or they are not serious in rating.

4. People can unduly influence the rating to favor their own work, or work by their friends, relatives or co-workers.

5. Ratings may not be set by people with the same values and views. For example, an expert in an area may prefer other choices than beginners. A resource which experts give bad ratings to, may be good for beginners. Also people values may influence their choices, for example a religious person may have different preferences than a cynical/sophisticated "modern" person.

Design of rating systems which better handle one of the above requirements may be less good for other requirements. For example, restricted selecting of who may provide the ratings may give higher-quality ratings (at least if the designer's values and views are the same as of those providing the rating) but reduce the amount of ratings and rated resources available.

To select only certain people who are allowed to provide ratings, or to let anyone provide ratings, but base selections on ratings made by people with the designer's values and views, are two alternative methods of getting higher-quality ratings. Is it an advantage to combine both methods, or will they interact so that one method is better than the other?

2.5 Collaborative Filtering Based Recommender Systems [4]

Recommender systems apply data analysis techniques to the problem of helping users find the items they would like to purchase at E-Commerce sites by producing a predicted likeliness score or a list of *top-N* recommended items for a given user. Item recommendations can be made using different methods. Recommendations can be based on demographics of the users, overall top selling items, or past buying habit of users as a predictor of future items. Collaborative Filtering (CF) is the most successful recommendation technique to date. The basic idea of CF-based algorithms is to provide item recommendations or predictions based on the opinions of other like-minded users. The opinions of users can be obtained explicitly from the users or by using some implicit measures. These opinions also called ratings.

2.5.1 User-based Collaborative Filtering Algorithms [6]

User-based algorithms utilize the entire user-item database to generate a prediction. These systems employ statistical techniques to find a set of users, known as *neighbors*, that have a history of agreeing with the target user (i.e., they either rate different items similarly or they tend to buy similar sets of items or they have similar demographic information). Once a neighborhood of users is formed, these systems use different algorithms to combine the preferences of neighbors to produce a prediction or *top-N* recommendation for the active user. The techniques, also known as *nearest-neighbor* or

customer-based collaborative filtering are more popular and widely used in practice.

2.5.2 Item-based Collaborative Filtering Algorithm [5]

Unlike the user-based collaborative filtering algorithm discussed above, the item-based approach looks into the set of items the target user has rated and computes how similar they are to the target item *i* and then selects *k* most similar items $\{i_1, i_2, \ldots, I_k\}$.

At the same time their corresponding similarities $\{s_{i1}, s_{i2}, \ldots, s_{ik}\}$ are also computed. Once the most similar items are found, the prediction is then computed by taking a weighted average of the target user's ratings on these similar items.

Chapter Four

Conclusion, and Future Work

4.1 Introduction

In this chapter, the conclusions of this work are given with some recommendations for future work.

4.2 Conclusion

From the results and the explanation of each method mentioned in chapter three, we conclude that a method accuracy depends on several factors:

1. The first factor is the speed of that method, because the customer prefers to spend less time waiting for the recommended items list.
2. The accuracy is different from customer to another. To explain this point, consider a customer rated an item and recommended using a particular –item based method with a list of items that satisfied him (her). And consider another customer who rated the same item, and recommended by the same item based method. The recommended items list will be definitely the same recommended list for the first customer, but the second one did not like this list. This will not mean that this method is more accurate for the first customer than for the second one. So, an important factor is the customers' opinions and desires. That's why the item-based1 method, that is powerful, had less accuracy than the others.
3. Another important factor is psychological conditions of the recommender system users. This means that if a user is satisfied with the recommendation list in one condition, it does not necessarily mean that the same user will be satisfied with the same recommendation list, but in another condition.

4.3 Future Work

There are many suggestions for future work related to this thesis subject. These suggestions are:

1- Improve the searching techniques used and enhance the speed of them.

2- It is a good idea to improve the system to deal with distributed database on many computers connected via a network.

3- The database can be expanded to contain further information like actors, authors, or directors' names.

4- Improving the system to recommend several kinds of items not only one kind.

5- One important point when using Customer-Based recommendation method is to use more demographic information about customers.

6- The assistant of a psychologist will be very important.

Chapter Three

Design and Analysis of E-Commerce Recommender System for Movies

3.1 Introduction

People face the problem of information overload every day. As the number of web sites, books, magazines, research papers, and so on continue to rise, it is getting harder to keep up. In recent years, recommender systems have emerged to help people find relevant information.

3.2 Recommendation Techniques

There are a number of techniques used for the design of a recommender system. We can categorize these techniques as Content-Based, and Social or Collaborative filtering (CF).

In content-based techniques, the user model includes information about the content of items of interest-- whether these are web pages, movies, music, or anything else. Using these items as a basis, the technique identifies similar items that are returned as recommendations. These techniques might prove highly suitable for users who have specific interests and who are looking for related recommendations. Many machine learning techniques have been applied to this problem. Some researchers working on these have modeled users with the application of neural network methodology.

One of the limitations when using content-based techniques is that no new topics are explored; only those that are similar to topics already in the user's profile. This leads to over-specialization: one is restricted to seeing items similar to those that have already been rated highly. This has been addressed in some cases with the injection of randomness. Content-based techniques, moreover, are difficult to apply to situations where the desirability of an item, for example a web page, is determined in part by multimedia content or aesthetic qualities. These types of materials are generally incompatible with the type of content analysis that these techniques require in order to make further recommendations.

Additionally, many recommender systems of this kind frequently require feedback about the relevance of their suggestions. Users often find generating this feedback a tedious task and try to avoid it. The user model in such systems consists entirely of user ratings of items of interest. Recommendations are solely based on these, making them the main factor influencing performance: the fewer the ratings, the more limited the set of possible recommendations. Feedback is required in machine learning techniques that need it for their "learning" process. These techniques often require lengthy computation to learn the user's preferences. Once computed, however, the user's preferences will not remain static. Therefore, this process will need to be repeated with a frequency that depends on how quickly the user model changes. [23]

The term "Collaborative Filtering" was explained by Paul Resnick1 who proposed the following working definition: "Guiding people's choices of what to read, what to look at, what to watch, what to listen to (the filtering part). And doing that guidance based on information gathered from some other people (the collaborative part)." [19]

[1] Paul Resnick is a professor at the University of Michigan's School of Information. He's worked extensively with recommender systems.

Collaborative filtering (CF) systems predict a person's affinity for items or information by connecting that person's recorded interests with the recorded interests of a community of people and sharing ratings between likeminded persons. [17]

Finding the "nearest neighbors" to the active user in order to retrieve recommendations is a task that requires the definition of the term "similarity" for a particular system. This is one of the main points where collaborative systems differ. Specifying which users are to be considered similar determines the performance of the system in terms of accuracy of recommendations. Keeping this in mind, a user that is considered unusual based on his profile (interests) will probably not be similar to any of the other users, which will lead to poor recommendations. Moreover, since no information about the content of items is kept, even users with similar (but not identical) interests will not be considered similar themselves.

The first collaborative filtering system was Tapestry and since then there has been significant research in the field. Several algorithms have been used for collaborative filtering, and specifically for computing the aforementioned similarity between two users.

The advantage of social (or collaborative) filtering, compared to content-based techniques, is that the pool from which recommendations originate is not restricted to items for which the active user has demonstrated interest. The pool will also include items that other users, users that are in some respect similar, have rated highly.

This can prove to be instrumental in enhancing the user's model: social filtering systems give the user the opportunity to explore new topics and items.

Breese et al. [8] divide the collaborative filtering algorithms into memory-based and model-based techniques. Memory-based collaborative filtering algorithms predict a user rating for a particular item by using a

similarity-weighted sum of the other user ratings. The method used to calculate weights is a distinguishing characteristic of algorithms in this category. Model-based algorithms depend on a model, such as a Bayesian network, built to represent the user data. This model will subsequently be queried to get the recommendations. The construction of the model is a learning process that is often time consuming.

Memory-based algorithms utilize the entire user-item database to generate a prediction. These systems employ statistical techniques to find a set of users, known as neighbors, that have a history of agreeing with the target user (i.e., they either rate different items similarly or they tend to buy similar set of items). Once a neighborhood of users is formed, these systems use different algorithms to combine the preferences of neighbors to produce a prediction or top-N recommendation for the active user. The techniques, also known as nearest-neighbor or user-based collaborative filtering are more popular and widely used in practice.

Model-based collaborative filtering algorithms provide item recommendation by first developing a model of user ratings. Algorithms in this category take a probabilistic approach and envision the collaborative filtering process as computing the expected value of a user prediction, given his/her ratings on other items. The model building process is performed by different machine learning algorithms such as Bayesian network, clustering, and rule-based approaches. [5]

Indicatively, the time required is very significant, especially when the user models are dynamic. However, the advantage here is that after the model is determined, recommendations can be returned with great speed.

In collaborative filtering, recommendations are often based on the comparison between the models of the active user and the population of other users, where the user models are sets of votes. A common

shortcoming of collaborative filtering algorithms recommendations will only come from the users with which the active user shares votes. [23]

A Collaborative Filtering algorithm should be both accurate (the recommended objects should subsequently receive high ratings), and efficient in terms of computational complexity. [18]

Social Information filtering exploits similarities between the tastes of different users to recommend (or advise against) items. It relies on the fact that people's tastes are not randomly distributed: there are general trends and patterns within the taste of a person and as well as between groups of people. Social Information filtering automates a process of ``word-of-mouth" recommendations. A significant difference is that instead of having to ask a couple friends about a few items, a social information filtering system can consider thousands of other people, and consider thousands of different items, all happening autonomously and automatically. The basic idea is:

1. The system maintains a *user profile*, a record of the user's interests (positive as well as negative) in specific items.
2. It compares this profile to the profiles of other users, and weighs each profile for its degree of similarity with the user's profile. The metric used to determine similarity can vary.
3. Finally, it considers a set of the most similar profiles, and uses information contained in them to recommend (or advise against) items to the user. [27]

Challenges of Collaborative Filtering Algorithms

Collaborative filtering systems have been very successful in past, but their widespread use has revealed some potential challenges such as:

• **Sparsely:** In practice, many commercial recommender systems are used to evaluate large item sets (e.g., **Amazon.com** recommends books

also **CDnow.com** recommends music albums). In these systems, even active users may have purchased well under 1% of the items (1% of 2 million books is 20, 000 books). Accordingly, a recommender system based on nearest neighbor algorithms may be unable to make any item recommendations for a particular user. As a result the accuracy of recommendations may be poor.

• **Scalability:** Nearest neighbor algorithms require computation that grows with both the number of users and the number of items. With millions of users and items, a typical web-based recommender system running existing algorithms will suffer serious scalability problems. [5]

3.3 The Recommendation Methods:

Hundreds of variants of algorithms related to the recommendation process have been published. Here, we will discuss Item-Based, Customer-Based and Intersection methods.

3.3.1 Item Based Method (Content-Based):

In content-based techniques (sometimes called item-to-item), the user model includes information about the content of items of interest--whether these are web pages, movies, music, or anything else. Using these items as a basis, the technique identifies similar items that are returned as recommendations. These techniques might prove highly suitable for users who have specific interests and who are looking for related recommendations. Many machine learning techniques have been applied to this problem.

Item-to-item collaborative filtering matches each of the user's purchased and rated items to similar items, then combines those similar items into a recommendation list. To determine the most-similar match for a given item, the algorithm builds a similar-items table by finding

items that customers tend to purchase together. We could build a product-to-product matrix by iterating through all item pairs and computing a similarity metric for each pair.

Algorithm 3.1 - Item-Based1 - Recommends the top N similar items to a single product by calculating similarities between all items:

Input: CAT, is the product catalog
I1, item in CAT
Pur_tab, is the purchasing table
C, is a customer in Pur_tab
I2, item in CAT

Output: Selection, is the table containing customer and two items bought by that customer
S, is a vector contains the items similarities to an item
Rec, is the recommendation list

```
For each item I1 in CAT
    For each customer C  in Pur_tab
        If C purchased I1 then
            For each item I2 in CAT
                If C purchased I2 and I1, and I1 is not equal
                to I2 then
                    Add to Selection that C purchased I1
                    and I2
            End for
    End for
    For each item I2 in Selection
        S(I2)= compute similarity between I1 an I2
    End for
    Sort S descending
    Put the top N items from S in Rec
En for
```

The result will be the vector Rec which contains the top N items that will be recommended to the user.

It's possible to compute the similarity (S) between two items in various ways, but a common method is to use the cosine measure [11]

$$Similarity(\vec{A},\vec{B}) = \cos(\theta) = \frac{\vec{A}\cdot\vec{B}}{\left\|\vec{A}\right\| * \left\|\vec{B}\right\|}$$

Where θ is the angle between A and B, and A and B are two vectors each one corresponds to an item rather than a customer, and the vector's M dimensions correspond to customers who have purchased that item.

The purchasing table Pure_tab contains customer, and items bought by that customer, i.e., if a customer (say 11) purchased items 4,7, then in the purchasing table it is presented like this:

Customer	Item
11	4
11	7

The algorithm works by trying to get the top N similar items to the item rated by the user, this is done by recording the items that previously bought by another customers with the item rated by the user. Then compute the most similar items among them to the item rated by the user. This similarity computation is done using the cosine measure. Then, recommend the topN similar items to the user. And then recommend these most similar items to the user.

One of the ideas in content-based algorithms works by asking the active user about the item he likes. Then compute the similarity between this item and the other items.

The top-N similar items to the rated item will be recommended to the active user.

The offline computation of the similar-items table (computing the similarity measure) is extremely time intensive, with $O(N^2M)$ as worst case where M is the number of customers and N is the number of product catalog items. In practice, however, it's closer to O(NM), as most customers have very few purchases.

Given a similar-items table, the algorithm finds items similar to each of the user's purchases and ratings, aggregates those items, and then recommends the most popular or correlated items. This computation is very quick, depending only on the number of items the user purchased or rated. We will refer to this method as Item-Based1.

Another way to recommend using the content-based technique (which was suggested during the work on the proposed system) is to recommend the items that are most related to the item bought by the active customer. We will refer to this method as Item-Based2.

This idea works as follows: it simply calculates the number of occurrences that the item rated by the active user bought with each item. Then take a list of items that have the top numbers of occurrences with the item rated by the active user. Select from this list top-N items and recommend it to the active user. Nearly, It is similar to Item-Based1, but it is different in similarity computation.

The output matrix Buy contains the number occurrences of all items with each item. To explain the matrix building well, let us have four items, study the following matrix that represents the Buy matrix:

	1	2	3	4
1	57	20	13	2
2	20	22	0	1
3	13	0	55	33
4	2	1	33	50

Figure 3.1 Example illustrates the Buy matrix

The cell Buy (4,1)=2 means that the items 4 and 1 purchased together twice. And Buy (1,2)=20 means that the items 1 and 2 purchased together 20 times. Where Buy (2,3)=0 means that items 2 and 3 never purchased together and so on. The shadowed cells (which is the diagonal of the matrix) represent the times of purchasing an item. So, Buy (2,2) represents the times of purchasing item 2.

Obviously, Buy (1,2)= Buy (2,1), and Buy (1,3)= Buy (3,1) and so on for the whole matrix. So, we do not need to calculate the triangle that lies below the diagonal.

After this table is created, for a particular customer, just compute the most related items to those the customer bought.

Algorithm 3.2 - Item-Based2 - Recommends the top N similar items to a single product by calculating the number of occurrences of each two items together:

Input: Pur_tab, is the purchasing table
I1, I2, items in Pur_tab
C, is a customer in Pur_tab

Output: Buy, is an N X N matrix holds the number of occurrences of each two items together
Rec, is the recommendation list

```
For each item I1 in Pur_tab
    For each item I2 in Pur_tab
        Buy(I1,I2)=0
    End for
End for
For each item I1 in Pur_tab
    For each customer C  in Pur_tab
        If C bought I1 then
            For each item I2 (not equal to I1) in Pur_tab
            bought by C
                Increment Buy(I1,I2)
            End for
    End for
    Put the top N occurred items with I1 in Rec
End for
```

The result will be the vector Rec which contains the top N items that will be recommended to the user.

This method is less time expensive than the method discussed earlier, since we do not need to compute the similarity equation.

However, many product pairs have no common customers, also computing the number of occurrences of an item with itself is useless, and thus the approach is inefficient in terms of processing memory usage.

3.3.2 Customer-Based Method (Collaborative Filtering)

In social or collaborative filtering, the system constructs rating profiles of its users, locates other users with similar rating profiles and returns items that the similar users rated highly. As in content-based techniques, these systems depend on their users providing ratings. [23]

The fundamental assumption is that if users A and B are similar, they share similar tastes, and hence will rate other items similarly. Approaches differ in how they define a “rating,” and how they define “similarly.”

Similarity is different from one recommender system to another. Some systems define customers A and B similar if they share the same rated items, some consider them similar if they have the same demographic information, and some say they are similar if they have some similar liked items.

In our proposed system, we considered customers A and B to be the similar if they have nearly similar demographic information, i.e., A is a user demanded recommendations, our method makes a list of the most similar customers to that user. Then the most similar customers to the active user (A) who will have the highest Sim values, we take the items that these customers bought which we call the recommendation list and recommend this list to the active user.

In the proposed system, customer-based algorithm supposes a matrix (Characteristic) in which number of rows represents the number of demographic information available in the database, and number of columns represents the number of customers in the database. Figure 3.2 is an example represents the matrix Characteristic.

In this example, there are three demographic-information: age (1), gender (2), and occupation (3), for example. We have also four customers, so, Characteristic (1,4)=57 means that the age of customer 4 equals 57, and Characteristic (2,4)=M means that the gender of customer 4 is male, and so on.

	1	2	3	4
1	33	20	13	57
2	M	F	M	M
3	Teacher	Student	Student	Doctor

Figure 3.2 Example illustrates the Characteristic matrix

Algorithm 3.3 - Customer-Based - Recommends the topN items depending on customers similarities to the active user:

Input: Cu_list, is the customers table
A, is the active user in Cu_list
i, is an index
Char_Num, is the number of characteristics available in the system
Characteristic, is the matrix of the demographic information of all the customers
Pur_tab, is the purchasing table
I1, is an item in the Pur_tab

Output: Sim, is a similarities vector of customers to the active user
C_Rec, is the most similar customers list
Rec, is the recommendation list

```
For each customer, C in  Cu_list
    Sim (C)=0
End for
For each customer in  Cu_list, C
    If C is not equal to A then
        For i=1 to Char_Num
            If Characteristic (i,A)= Characteristic (i,C) then
            Increment Sim(C)
End for
Put the customers with the top N Sim values in C_Rec
For each customer C in C_Rec
    For each item I1 in Pur_tab and  purchased by C
        Add I1 to Rec
    End for
End for
```

The result will be the vector Rec which contains the top N items that will be recommended to the user.

The demographic information is a basic measure. People from the same age or the same social layer highly share the same tastes. Also the psychological situation of each person in a particular time can determine the person's decisions. A psychological analysis is an interesting point that could be taken in consideration in building recommender systems in the future.

This method is less time expensive and even memory storage than the item-based. But an important point is that if there much demographic information available about customers, the algorithm works well, but if they are few, the accuracy will be less.

3.3.3 Customer and Item Based Method (Intersection)

This method is invented during work on the proposed system to give it more power. It depends on item-based and customer-based methods. First, we must have two lists of items recommended by Item-based and customer-based, then we take the intersection of these two lists. The result will be the items to recommend by the method "Intersection". This method is powerful. But, it requires execution of two algorithms (Item-based and Customer-based). So, it will definitely be the most expensive algorithm in time.

This method can enhance and strength the recommendation, but also it can give it weakness when the intersection between the item-based and the customer-based is empty.

Because Item-based2 is faster than Item-based1 (see Item-Based2 Method), using it with Item-based1 in the proposed system is better.

3.4 Experimental Data

We performed experiments on a subset of movie rating data collected from the MovieLens web-based recommender (movielens.umn.edu) designed by The GroupLens Research Project at the University of Minnesota.

The proposed system is used to guide the user through different methods by recommending items that will most probably suite the user interests. Recommending these items is a prediction based on information taken from the user profile.

As discussed in 2.4.1 before, recommendation operation includes five steps; selection, preprocessing and transformation, data mining, interpretation and evaluating, and presentation.

The GroupLens selected the items and the ratings on these items, then preprocess and transform the data, after that used data mining methods. We used the resulted database in our system. So, we started from fourth step (Interpretation and evaluating).

3.5 The Experimental Database

In the experimental database, there are 943 users and 1682 movies. Each user rated at least for 20 movies. There is simple demographic information about the users. These are age, gender, occupation, and Zip. The users rated for the films from 1 to 5. These degrees are:

5 Excellent

4 Very Good

3 Good

2 Medium

1 Bad

The experimental database consists of the following tables:

1. Movies info: Contains full information about movies. It has the following columns names: Movie Id, Movie Title, Release Date, Video Release, while The other 19 columns are: unknown, Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western. These fields are the genres, a 1 indicates the movie is of that genre, a 0 indicates it is not; movies can be in several genres at once, for example, a movie can be a romance and action at the same time.
2. Ratings: Has the following columns: id, movie id, rating.
3. Userinfo: With the following columns: id, age, gender, occupation, and zip.

3.6 The Proposed System Interface Structure:

The block diagram of the proposed system is shown in figure 3.3.

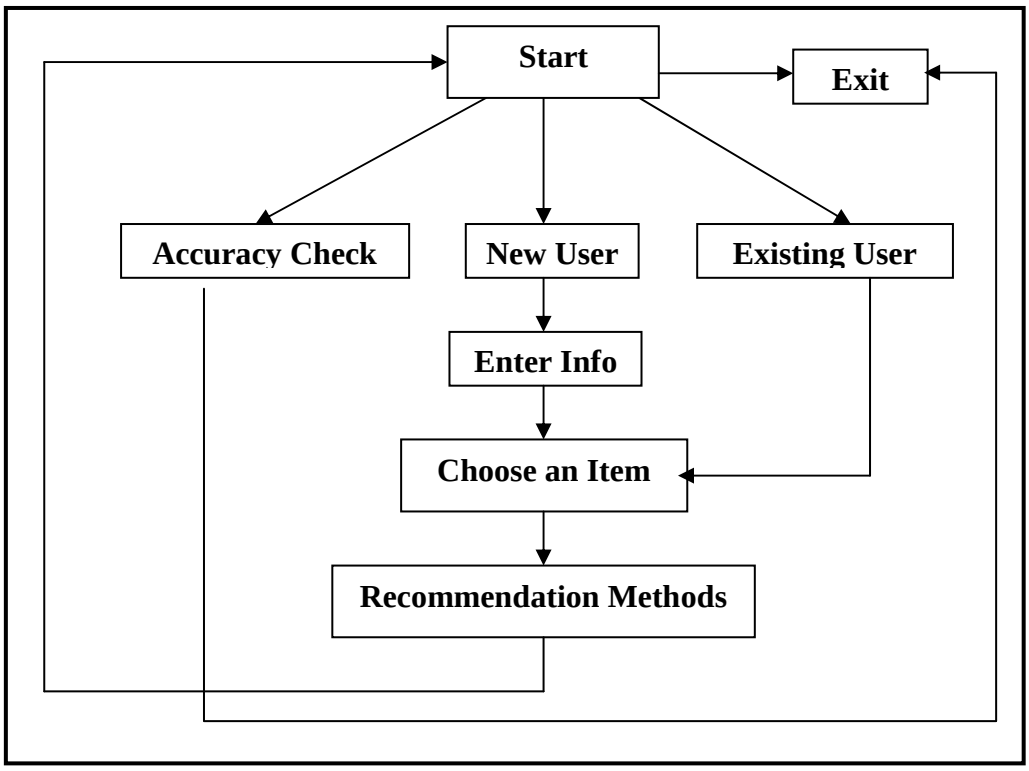

Figure 3.3: The Proposed System Interface Structure

The first screen appears is the “Start” from which the user goes to one of the three options: Accuracy Check, New User, or Existing User.

If this is the first time for the user to use this system, he/she goes to New User then he will be asked for personal information (age, gender, and occupation). In this case the system automatically gives the user an id.

Else, if the user was an existed user in the database he will choose to go to Existing User and gives his id.

Both of the New User and Existing User need to choose an item from the list of items that are sorted according to high ratings. Then the user chooses a recommendation method from the four methods available

in the system (Item-Based1, Item-based2, Customer-based, and Intersection).

A list of recommended items will appear to the user according to the method chosen by him.

If the user wanted to check the accuracy of a method, he just head to the Accuracy Check and select the method, then just wait for the accuracy to be calculated.

3.7 Results and Discussion

To test the proposed system methods' accuracy, we used two ways. In the first way, we simply selected real users in the database that already bought movies and recommend them using our methods. The accuracy will be computed as percentage of the similarity between the results and what they really bought.

We selected first customer number 1 who supposed that he rated to movie number 1, then we recommend him using our four methods. Then we compute the similarity between the items he really bought and the items we recommended for each method. Then we selected customer number 7 who supposed to rate to movie number 10 and also the accuracy was computed for the four methods. Table 3.1 shows the accuracy for each method for the two customers.

	Customer No.1	Customer No.7
Item-Based1 Method	94.47 %	94.62 %
Item-Based2 Method	99.57 %	99.43 %
Customer-Based Method	54.47 %	91.50 %
Intersection	5.532 %	9.915 %

Table 3.1 Methods Accuracy for Customers 1 and 7.

In the second way, we presented the list of items to a sample of my colleagues and asked them to rate for one item, then they recommended using our four methods depending on the items they chose and their personal information. Then we asked them to give a degree in the range 1-100 to each method. The degree of a particular method was considered to be the accuracy of that method from the user's point of view who rated it.

To discuss the results, we should go back to table 3.1. In this table, we have eight cases, we will discuss each one.

- **Case1 (Customer1 recommended using Item-Based1 method) and Case2 (Customer7 recommended using Item-Based1 method):**

As we can see, the accuracy for the two cases is high (94.47, 94.62), this is because the similarity measure which gives Item-Based1 method more power.

- **Case3 (Customer1 recommended using Item-Based2 method) and Case4 (Customer7 recommended using Item-Based2 method):**

The accuracy for these two cases is high, which means the method is a good one.

When comparing cases 3 and 4 with the first two cases, we can see that the accuracy of Item-Based2 is higher than the accuracy of Item-based1.

The method accuracy is different from time to time depending on the customer's demographic information, the items he/she rated, and the

number of these items. So, the results do not necessarily mean that Item-Based2 always more accurate than Item-Based1.

- **Case5 (Customer1 recommended using Customer-Based method) and Case6 (Customer7 recommended using Customer -Based method):**

Here we can see a large difference between cases 5 and 6. This difference happened because Customer-Based method depends on the customer demographic information which is obviously different from customers 1 and 7. Table 3.2 shows the demographic information for customers 1 and 7.

	Age	Gender	Occupation
Customer No.1	24	Male	Technician
Customer No.7	57	Male	Administrator

Table 3.2 The demographic information for customers 1 and 7

- **Case7 (Customer1 recommended using Intersection method) and Case8 (Customer7 recommended using Intersection method):**

As we can see the accuracy in these two cases is less than the other cases. As discussed before in chapter three, we compare the resulted recommendation list to the real bought items by the customers. Intersection reduces the items by selecting the similar items in two lists. This also reduces the accuracy, because the reduced list will be less similar to the real bought list.

Recommender system

Recommendation methods

E-commerce

Data mining

Rating

Collaborative filtering

User based

Item based

Internet

Web site

List of Contents

الاسم	أسيل باسم صبري الطائي
تاريخ المناقشة	٢٨-١١-٢٠٠٤
الهاتف	5223525, 07901887322
العنوان	بغداد- الحرية – م ٤١٨ – ز ٢٥ – د٢٧
الايميل	aseel_basim@yahoo.com

References

1. Adams Site, “What Is E-Commerce?“ http://www.adamssite.com/what-is-e-commerce.shtml
2. Mr. Ameen Damani, “Description and Evaluation of Different Types of E-Payment Systems” http://www.witiger.com/ecommerce/paymentmatrix.htm
3. Andreas Geyer-Schulz and Michael Hahsler, “Evaluation of Recommender Algorithms for an Internet Information Broker based on Simple Association Rules and on the Repeat-Buying Theory” http://wwwai.wu-wien.ac.at/~hahsler/research/recomm_webkdd2002/final/webkdd2002.pdf
4. Badrul M. Sarwar, “Collaborative Filtering Based Recommender Systems”, February 19, 2001 http://www.unizh.ch/home/mazzo/reports/www10conf/papers/519/node5.html
5. Badrul M. Sarwar, “Item-Based Collaborative Filtering Algorithm”, February 19, 2001 http://decweb.ethz.ch/WWW10/papers/519/node10.html
6. Badrul M. Sarwar, “Memory-based Collaborative Filtering Algorithms”, February 19, 2001 http://decweb.ethz.ch/WWW10/papers/519/node7.html
7. Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, “Item-based Collaborative Filtering Recommendation Algorithms”, GroupLens Research Group/Army HPC Research Center/ Department of Computer Science and Engineering/ University

of Minnesota. May 1-5, 2001, Hong Kong. http://www-users.cs.umn.edu/~karypis/publications/Papers/PDF/www10_sarwar.pdf

8. Breese, Heckermen, and Kadie, “Empirical analysis of predictive algorithms for collaborative filtering”, Microsoft Research Technical Report, (MSR-TR-98-12), October 1998.
9. Byung-Kwan Lee, “New Technology and its Impact on Online Data Collection” http://www.ciadvertising.org/student_account/fall_00/adv391k/bklee/paper/frame-tech1.html
10. Cabena, Hadjinian, Stadler, Verhees, Zanasi, “Discovering data mining from concept to implementation”, Prentice Hall. 1997.
11. Greg Linden, Brent Smith, and Jeremy York January, “Amazon.com Recommendations/ Item-to-Item Collaborative Filtering”, February 2003 http://www2.bc.edu/~heimgr/md254s03/IEEEInternetComputing_AmazonComRecommendationsCollabFiltering_w1076.pdf
12. HM Customs and Excise, “What is E-Commerce?” http://www.hmce.gov.uk/business/tradinginternet/tradinter-3.htm
13. Jacob Palme, “Select EU-funded Project” http://cmc.dsv.su.se/select/rating-choices.html
14. Jamal Adnan Asaad, “Using Data Mining For Decision Support”, M.Sc. Thesis Submitted To The Iraqi Commission For Computers And Informatics/Informatics Institute For Postgraduate Studies, Baghdad 2002
15. J. Ben Schafer, Joseph A. Konstan, John Riedl, “E-Commerce Recommendation Applications”, GroupLens Research Project/ Department of Computer Science and Engineering- University of

Minnesota.
http://www.cs.umn.edu/Research/GroupLens/papers/pdf/ECRA.pdf

16. Joan Silvi, “Recommender Systems”, February 6 and 7, 1999 http://www.iota.org/Winter99/recommend.html

17. Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl, “Explaining Collaborative Filtering Recommendations”, Dept. of Computer Science and Engineering University of Minnesota Minneapolis, MN 55112 USA http://web.engr.oregonstate.edu/~herlock/papers/explanations_cscw2000.pdf

18. Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins, “Eigentaste: A Constant Time Collaborative Filtering Algorithm”, from IEOR and EECS Departments University of California, Berkeley August 2000 http://www.ieor.berkeley.edu/~goldberg/pubs/eigentaste.pdf

19. Louise A. Arnheim, “COLLABORATIVE FILTERING WORKSHOP”, for the Coalition for Networked Information (CNI). March 16, 1996 -- Berkeley, CA http://www.sims.berkeley.edu/resources/collab/collab-report.html

20. Mukund Deshpande and George Karypis, “Selective Markov Models for Predicting Web-Page Accesses”, University of Minnesota, Department of Computer Science/Army HPC Research Center http://www.siam.org/meetings/SDM01/pdf/sdm01_04.pdf

21. Norbert Turek, “E-Commerce Counts On Servers-Choosing the right commerce-server vendor is a critical decision”, September 13, 1999 http://www.iweek.com/752/servers.htm

22. Osmar R. Zaïane, “Principles of Knowledge Discovery in Databases”, 1999 http://www.exinfm.com/pdffiles/intro_dm.pdf

23. Patrick Paulson1 and Aimilia Tzanavari, “Combining Collaborative and Content-Based Filtering Using Conceptual Graphs”, Miami University Department of Computer Science & Systems Analysis. http://www2.cs.ucy.ac.cy/~aimilia/pubs/Paulson-Tzanavari-revised.pdf
24. Paul Resnick and Hal R. Varian, “Recommender Systems”, Guest Editors 1997 http://www.acm.org/pubs/cacm/MAR97/resnick.html
25. Rachel Konrad, “Data mining: Digging user info for gold Data mining makes inroads”, February 9, 2001 http://www.techupdate.com/techupdate/stories/main/0%2C14179%2C2683567-4%2C00.html
26. Supiya Ujjin and Peter J. Bentley, “Building a Lifestyle Recommender System”, University College London/ Department of Computer Science http://www10.org/cdrom/posters/1039.pdf
27. Upendra Shardanand and Pattie Maes MIT Media-Lab, “Social Information Filtering: Algorithms for Automating -Word of Mouth-“, Cambridge http://www.acm.org/sigchi/chi95/Electronic/documnts/papers/us_bdy.htm
28. “A Brief History of the Internet and Related Networks” http://www.pradeepkumar.20m.com/history.htm
29. “Background Of Electronic Commerce” http://www.personal.psu.edu/users/k/g/kgg113/spcom/groupweb/ecom.html
30. “Data Mining” http://www.cs.aue.auc.dk/datamining/papers/yijunlumsc.ps
31. “What Is E-Commerce” Swainsboro Web Hosting. http://www.swainsborowebhosting.com/e-Commerce.html

32. “What Is E-Commerce- Introduction” NWT 2000
http://www.media-miracles.co.uk/html/ecommerce.htm

Republic of Iraq
Ministry of Higher Education
Al-Nahrain University
College of science

Recommender System for E-Commerce Data

A Thesis Submitted To The
College Of Science, Al-Nahrain University
In Partial Fulfillment Of The Requirements
For
The Degree Of Master Of Science In
Computer Science

By
Aseel Basim Sabry Yakoob Al-Tayee
(B.Sc. 2001)

Supervisor
Dr. Taha Saadoon Bashaga

August 2004 **Jamada Al-Thani 1425**

جمهورية العراق
وزارة التعليم العالي
جامعة النهرين
كلية العلوم

النظام الناصح للتجارة الالكترونية

رسالة مقدمة إلى كلية العلوم، جامعة النهرين كجزء من متطلبات نيل شهادة الماجستير في علوم الحاسبات

من قبل

أسيل باسم صبري يعقوب الطائي

بكالوريوس

٢٠٠١

المشرف

د. طه سعدون باشاغا

آب ٢٠٠٤ جمادى الثاني ١٤٢٥