INTERNET BASED MOVIE GENRE SUGGESTION MODEL CONSIDERING DEMOGRAPHICAL INFORMATION OF USERS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

TUNA HACALOĞLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

APRIL 2013

INTERNET BASED MOVIE GENRE SUGGESTION MODEL CONSIDERING DEMOGRAPHICAL INFORMATION OF USERS

Submitted by **Tuna Hacaloğlu** in partial fulfillment of the requirements for the degree of **Master of Science in Information Systems**, **Middle East Technical University** by,

Prof. Dr. Nazife Baykal ____________________
Director, **Informatics Institute**

Prof. Dr. Yasemin Yardımcı Çetin ____________________
Head of Department, **Information Systems**

Assoc. Prof. Dr. Sevgi Özkan ____________________
Supervisor, **Information Systems, METU**

Examining Committee Members:

Assist. Prof. Dr. Erhan Eren ____________________
Information Systems, METU

Assoc. Prof. Dr. Sevgi Özkan ____________________
Information Systems, METU

Assist. Prof. Dr. Banu Günel ____________________
Information Systems, METU

Aylin Akça Okan ____________________
(ATOS Bil. ve Dan. A.Ş.)

Assist. Prof. Dr. Çiğdem Turhan ____________________
(Atılım University, Software Engineering)

Date: April 17, 2013

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Tuna Hacaloğlu

Signature: ________________

ABSTRACT

INTERNET BASED MOVIE GENRE SUGGESTION MODEL CONSIDERING DEMOGRAPHICAL INFORMATION OF USERS

HACALOĞLU, Tuna
M.S., Department of Information Systems
Supervisor: Assoc. Prof. Dr. Sevgi ÖZKAN

April 2013, 86 pages

Web based customer recommendation systems are being used to provide customized information to the users. They are applied in many areas such as web browsing, information filtering, news or movie recommendation, and e-commerce. The primary aim is to offer suggestions about products or services that users might be interested in. They are intelligent applications to assist users in a decision-making process where they want to choose one item amongst a potentially large set of alternative products or services. These systems are based on information filtering. There are various types of information filtering methods that are used in these systems such as collaborative filtering, content-based filtering and hybrid methods. These types diverge according to the data that they focus on. For example some of them focus on finding similar items where others focus on similar customers. The key component of all recommendation systems is the user model which contains knowledge about the user's choices, preferences, and past activities which determine his behavior, in other words, his activities on the web. The recommendation systems working mechanism can be summarized in two steps: user model construction and recommendation generation. In this study, a prediction method is proposed according to the structure of the customer spectrum. Considering demographic data of users such as gender, age, education and occupation, the movie genre choice of the users is predicted. A comparison of two different methods will be given in the study on the online raw data provided by an online shopping site.

Keywords: Naive Bayesian Classifier Decision Tree, Classification and Regression Trees (CART), Chi-squared Automatic Interaction Detector (CHAID), Recommender Systems

ÖZ

KULLANICILARIN DEMOGRAFİK BİLGİLERİNE DAYALI İNTERNET TABANLI FİLM TURU ÖNERİ MODELİ

HACALOĞLU, TUNA
Yüksek Lisans, Bilişim Sistemleri
Tez Yöneticisi: Doç. Dr. Sevgi ÖZKAN

Nisan 2013, 86 sayfa

Web tabanlı müşteri tahminleme sistemleri kullanıcıya özel bilgi sağlamak için kullanılmaktadır. Bu sistemler web tarama, bilgi filtreleme, haber veya film tavsiye ve e-ticaret gibi birçok alanda uygulanmaktadır. Bu sistemlerin temel amaçları kullanıcının ilgisini çekebilecek ürünler veya servisler ile ilgili öneriler sunmaktır. Tahminleme sistemleri, büyük bir alternatif ürün veya servis kümesi içinde kullanıcıya karar verme sürecinde yardımcı olan akıllı uygulamalardır. Bu sistemler temelde veri filtrelemeye dayanırlar. Veri filtrelemenin bu sistemlerde kullanılan ortak filtreleme, içerik tabanlı filtreleme ve karma yöntemler gibi çeşitli tipleri mevcuttur. Bu tipler odaklandıkları veriye gore birbirlerinden ayrılmaktadırlar. Örneğin bazı yöntemler benzer öğeleri ele alırken, diğerleri benzer kullanıcıları ele almaktadırlar. Tahminleme sistemlerinin anahtar elemanı kullanıcının seçimleri, tercihleri, ilgi alanları ve geçmiş etkinliklerine dair bilgi içeren bir kullanıcı modelidir. Tahminleme sistemlerinin çalışma mekanizması iki adımda özetlenebilir:

Kullanıcı modeli oluşturulması ve tahminleme üretimi. Bu çalışmada, geniş bir müşteri yelpazesine sahip bir e-alışveriş sitesi için kullanıcıların film tercihlerini tahminleme ele alınacaktır. Müşteri bilgilerinin yapısına göre başka bir deyişle, müşterilerin cinsiyet, yaş, eğitim, meslek vb. demografik bilgilerini kullanarak bir film türü öneri modeli bu çalışmanın kapsamını oluşturmaktadır. Ayrıca çalışmada iki farklı metodun karşılaştırması online ham veri üzerinden verilecektir.

Anahtar Kelimeler:Naive Bayes Sınıflandırıcı, Karar Ağaçları, CHAID analizi, CART analizi, Öneri Sistemleri

To the memories of my grand mother

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor *Assoc. Prof. Dr. Sevgi OZKAN* for her guidance, sharing of her knowledge, and understanding during this study. She encouraged me in each step from the beginning to the end of this study. I am grateful to her.

I would like to express my deepest thanks to *Aylin AKCA OKAN* for her invaluable help and support during this study both motivating by trusting in me and sharing her profound knowledge with me. I will always be grateful to her.

I would like to express special thanks to *Assist. Prof. Dr. Deepti MISHRA* for her encouragement, help, trust, and patience. Her suggestions always motivated me and helped me to find the best way.

I would like to thank to *Oya Deniz BEYAN* for her encouragement and help.

I would like to thank to my mother *Nazan HACALOGLU,* my father *Tuncay HACALOGLU* and my sister *Tugce HACALOGLU* for their invaluable love and support during my life.

I want to especially thank to my dear friends who were the witnesses of this stressful period and who alleviated this stressful load with their continuous encouragement and friendship.

I would like to thank to my teachers in the *Departments of Computer & Software & Information Systems Engineering of ATILIM UNIVERSITY* for their guidance and collaborations.

I would like to express thanks to the thesis committee, *Assist. Prof. Dr. Erhan EREN, Assist. Prof. Dr. Banu GUNEL, Assist. Prof. Dr. Cigdem TURHAN,* and *Aylin AKCA OKAN* for their valuable suggestions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

NBC	Naïve Bayes Classifier
DT	Decision Tree
CHAID	Chi-Squared Automatic Interaction Detection
CART	Classification and Regression Tree
CF	Collaborative Filtering
CBF	Content-based Filtering

CHAPTER 1

INTRODUCTION

The worldwide spread of the internet and the tremendous continuing growth of the contents supplied by the web caused a need for applications which help users to choose the product that is related to them in the web. There is a fact that web contains an infinite number of information and from this large information pool, users - by eliminating the information that is irrelevant to them - should find the relevant one. Particularly, with the improvements in e-commerce technologies, today, there are lots of e-stores serving to their customers through web. Especially, in the past several years, the change in the way of surfing the web can be easily perceptible. Earlier, the users were offered generic choices by the websites. These choices are the ones that were offered to everybody. In these types of systems, users were in a state in which they have a lot of alternatives to choose and need a guidance to diminish the whole set of alternatives into a meaningful set (Pazzani, 1999). Therefore, earlier users were adapted to the websites. However, now the websites have started to adapt to their users and the websites offer their users some choices. The choices are some items that can be commercial products like furniture, clothes; cultural and entertainment products like movies, books, TV programs, travel alternatives, advertisements, blog articles; or educational subjects such as academic papers. Here the key point is that, the website performs this activity by considering lots of particularities of their users such as demographic data, past activity, by looking similarity among other users, or by looking at the item attributes with the help of some intelligent applications. These applications are called recommender systems or recommendation systems in the literature and there are many examples of these systems that we are used on the web. Shih et al. (2011) define the recommender system as a system which recommends an item or service to users regarding users' needs. The main idea behind these systems is to provide personalized information to the users according to their interests.

The recommendations can be done according to the approach and the recommender systems adopt such as a system can suggest an item to a customer based on his past activity, or by combining him with another user and suggesting an item accordingly, or by giving recommendation according to the customer's specifications and so on. These methods constitute the major types of recommender systems. The main purpose of e-commerce companies for using recommender systems is the mainstream rule of commerce which is increasing the sales. Schafer et al. (1999) in their study explain that recommender systems promote e-commerce sales in three ways such as "browsers into buyers", "cross-sell" and "loyalty". It means that with the help of recommender systems, people who are just browsing the page become customer since personalized suggestions enable them to find items of interest. Similarly, according to the items that are already in the shopping cart, recommender systems offer new items of interest and this kind of interaction increases the customer loyalty which in return increases the sales of the company. Hung (2005) stated that "A personalized recommendation system can help enterprises launch one-to-one marketing" and therefore "good recommendation can boost sales which means that customers are willing to spend more money and continually buy more products."

1.1. Motivation and Background

Researchers broadly divide recommender systems into two categories as Collaborative Recommendation and Content-based filtering (Adomavicius & Tuzhilin, 2005). However, there are other categories as knowledge-based, demographic and hybrid recommender systems which have been studied by the researchers. The different types of recommender systems are described in more detail in Chapter 2, section 2.1. Despite most of the current research, improvements and comparative studies are based on collaborative, content-based and hybrid recommendation fields and these studies are mostly concentrated on user-item ratings, item attributes rather than user demographic data, we find noteworthy to consider demographic attributes of people in the recommendation process. The reason is that demographic data is generally a primary data that a user provides to an online system, through registration, and can be the only data to be used to offer a recommendation to users even if when no past behavior is available. Moreover, it is appropriate for the cases in which users' feedbacks are not easily collected. People are not willing to rate / rank items or do not show a proper buying in some websites therefore some recommendations should be provided to them, with the limited information. For this reason, in this study we approach to recommendation strategy from the demographic side. The movie domain is one of the most attractive domains for the researchers in which a lot of studies have been conducted. When the literature of the recommender system is investigated, it is seen that many of the studies either targeted to the movie domain or use data from the movie domain for their validation. According to a literature review conducted by Park et al. (2012) on the classification of recommender system studies, it is seen that movie domain is used in 53 papers over 210 research papers published between 2001 and 2010. It means that movie domain constitutes one-fourth of the application field of recommender system research. This fact shows that although movie domain is not a new area for the recommender system research, it still captures the attention of the researchers.

There are a lot of methods offered in the literature to improve the recommender system research. In this study, we adopt the Naïve Bayesian Classifier (NBC) and Decision Trees (DT) methods to create a recommender system. Primarily, users are assigned to the movie genre choices with respect to their demographic profile such as age, gender, education and occupation as well as their time and day preferences by using the Naïve Bayesian Classifier (NBC). Since NBC assumes that each attribute is independent of each other and has equal contribution to the result, it is suitable for the cases in which users do not provide all of their demographic details. However, this independence assumption is not always true in reality. Therefore, it is needed to eliminate the redundant attributes. When the literature is investigated, it is found that Decision Trees are quite good for the feature selection of Naïve Bayesian Classifier and Decision Tree enhanced NBC outperforms the original Naïve Bayesian Classifier in various studies (Ratanamahatana & Gunopulos, 2003; Hall, 2007; Gayatri et al., 2010). However, none of the previous studies investigated the performance of this combined technique neither in the movie domain by considering users' demographics, nor in the recommender systems field. It means that previous research did not study to the impact of DT enhanced NBC in the recommender systems area. Hence, for the feature selection of NBC, decision tree approach which has adopted in many previous studies in the literature is adopted in this study. Hence the motivation of this study is that previous researches did not test decision tree enhanced NBC in the movie domain. Moreover, to the best of our knowledge, no recommender system is studied with the combination of these techniques before. For this reason, first of all, two different decision tree algorithms are investigated for the feature selection of NBC and these algorithms are CHAID-Chi-Squared Automatic Interaction Detector and CART (also known as CART) - Classification and Regression Tree algorithms. With the help of CHAID decision tree technique, the demographic attributes which are correlated or which contain redundant information are eliminated and then NBC is applied to the remaining attributes. It is seen that, CHAID eliminates many of demographic attributes and therefore another decision tree algorithm CART is used to identify important attributes instead of CHAID.

1.2. Objectives and Contributions

To the best of our knowledge, no such personalized system exists in the literature that uses customers' demographic attributes to predict and recommend movie genre by the help of the Naive Bayesian Classifier and additionally using Decision Trees to select the most significant attributes in the process. This study has the following research contributions:

- Evaluation of NBC for the classification performance of users' demographics and day and time preferences to movie genres
- Presentation of the hybridization of NBC and CART decision tree techniques for the classification
- Development of a personalized movie genre recommender system prototype for movie genre recommendation using demographic particularities of users, users' day and time preferences and feedback.
- Validation of the model with 28 people in real data.

The prototype system is tested with the help of 28 people. The results of this experiment are used to measure the performance of the techniques with the help of accuracy, recall and precision metrics. It is found that decision trees are useful for feature selection of NBC in the movie domain and showed similar improvements like other studies performed in other domains. Moreover, a recommender system benefiting from both NBC and users' feedback showed a learning ability and improvement in the recommendations over time.

The stages of the study are given in the following Figure 1.

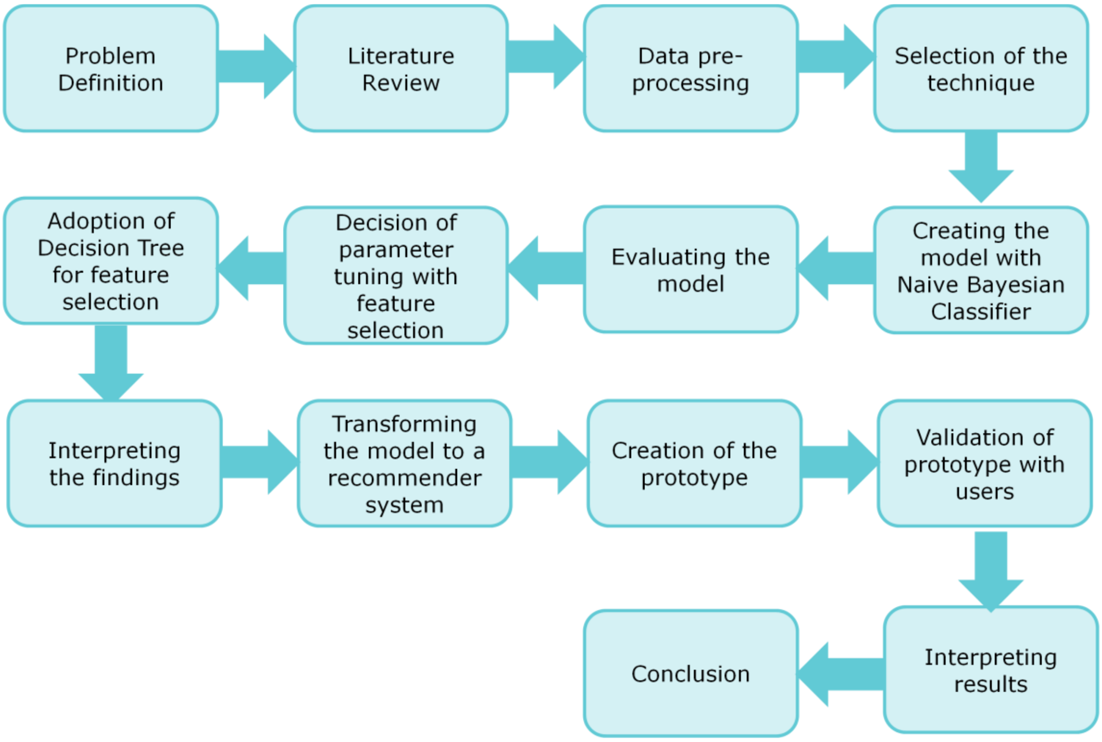

Problem Definition
Literature Review
Data pre-processing
Selection of the technique
Creating the model with Naive Bayesian Classifier
Evaluating the model
Decision of parameter tuning with feature selection
Adoption of Decision Tree for feature selection
Interpreting the findings
Transforming the model to a recommender system
Creation of the prototype
Validation of prototype with users
Interpreting results
Conclusion

Figure 1: Stages of the study

While conducting this study, some assumptions were needed to be taken into account. The key assumptions are presented in the following section 1.3.

1.3. Assumptions

One of the primary assumptions comes from the technique adopted in this study which is the NBC. NBC assumes that all the attributes are independent of each other. This is not generally the case in real life.

Another assumption is that movie genre preference is not a unique choice. It is not a correct argument that people who like drama do not prefer horror movies. They can prefer both. It is assumed that people may prefer three genre the most. In this specific study, people are offered recommendation that is not only the movie genre which has the maximum probability, but also the genres in the second and third level. For this reason, the result of NBC which classifies the instances to the class which has maximum probability is extended to the top three genres. In the literature, this fact is called as “second guess heuristic” (Han & Kamber, 2006) and used for the situations in which an instance can belong to more than a single class. It is stated that “A class prediction is judged as correct if it agrees with the first or second most probable class” (Han & Kamber, 2006).

In other respects, there are a lot of attributes that define a movie such as cast, director, genre, year, award, country, language, production year etc. However only, genre attribute of the movie is taken into consideration for this specific study to describe a movie. Additionally, a movie is not described by just one genre. It is quite common that a movie can be composed of more than one genre. A movie can be comedy-drama or romance-action-comedy at the same time. However, to decrease the number of movie genre classes and to increase the matching probability, it is assumed that a movie is described by only one genre.

1.4. Structure of the Thesis

The remaining chapters of this thesis are organized as follows: In Chapter 2, the review of the literature is given. Recommender systems and their different types are described along with their advantages and disadvantages. Later, the particularities of demographic recommender systems which are adopted in this study are given with significant studies performed in the literature. Next, recommender systems that adopted NBC are presented.

In Chapter 3, the methodology of this study is explained. The data description and data-processing steps are provided first. Later, NBC approach which is the core technique used in this study is explained. The DT algorithms such as CHAID and CART used for the feature selection are provided in detail. Additionally, the personalization of the recommendation approach is given in this section too.

In Chapter 4, the prototype system is described along with the system architecture implementation details; and explanations for the validation step are provided.

In Chapter 5, results of the experiment are presented along with the discussion.

In Chapter 6, conclusion and future research directions are presented.

CHAPTER 2

LITERATURE REVIEW

In this chapter, the review of the literature is introduced. Recommender systems and their different types are described. Later, particularly demographic recommendation approaches that are adopted in this study are presented. Next, data mining techniques used in recommender systems research field, specifically NBC and DT used in this study are provided.

2.1. Recommender Systems

Recommender Systems are the intelligent engines that are used to personalize the web content according to the customer preferences. The contents can be commercial products like furniture, clothes; cultural and entertainment products like movies, books, TV programs, travel alternatives, advertisements, blog articles; or educational subjects such as academic papers. Although there is a significant number of research conducted in the recommender systems field, the area is still immature. Many researchers still try to find the best approach for the recommendations. Recommender systems are classified into some categories with respect to their application techniques. Researchers are still not agreeing on the classification of recommender systems. Researchers divide the recommender systems generally into two categories: Collaborative Approaches and Content-based Approaches (Adomavicius & Tuzhilin, 2005). Shih et al. (2011) pointed out that the recommender systems have mainly two categories.

However, they also claim that hybrid approaches are also as much popular as other two approaches and the main recommendation approaches can be categorized as three namely collaborative approaches, content-based approaches and hybrid approaches. Stritt et al. (2007), Jannach et al. (2011) stated that Recommender Systems are divided into four categories as Collaborative Filtering, Content-based Filtering, Knowledge-based Recommendation and Hybrid Recommendations. Montanar et al. (2003), Sobecki (2006), Ghazanfar and Prugel-Bennett (2010) state that recommender systems have three types namely collaborative-filtering, content-based filtering and demographic recommender systems. Jannach et al. (2011) argue that Knowledge-based Recommendation and Hybrid Recommendations are other categories. Also, in some research, it is encountered that Demographic Recommendation also constitutes an integral part of the recommendation approaches. For example, Mahony et al. (2003) and Ricci et al. (2011) divide the recommendation approaches into six categories as content-based, collaborative, demographic, knowledge-based, and community-based and hybrid approaches. Recently, in the literature, there are many studies conducted in especially collaborative, content-based and hybrid approaches. The data that the recommender systems use in their algorithms are demographic data, rating data, behavior pattern data and transaction data (Wei et al., 2007). In the following section, collaborative filtering, content-based filtering, demographic filtering and hybrid methods will be presented.

2.1.1. Collaborative Filtering Recommender Systems

Collaborative Filtering (CF) is the first introduced recommendation technique in mid-1990s (Resnick et al., 1994; Shardanand & Maes, 1995; Herlocker et al., 2001) and it is the most used recommendation technique in the literature (Ahn 2008; Bobadilla et al., 2012). Herlocker et al. (2004) defines the CF as the most successful technique among the others. The main idea of CF is to investigate the relationship between two users who have similar tastes and recommend the item that one user has selected, observed, rated, bought, or purchased to other users by looking at some similarity measures. Thus, this technique aims to offer predictions of items to a user, by looking at other people having similar interests with the current user, with the help of statistical analysis of explicitly from the evaluations of users or implicitly by observing the behavior of the users (Montaner, 2003). In other words this technique correlates personal preferences of people. In CF, there is no need to keep data about the items because this technique looks at the user rating history to find a correlation between the current user and other users who share the same tastes. Collaborative filtering uses ratings to get the users preferences. Since it correlates different users, it can be said that it has a social side. Collaborative filtering methods can be categorized into two types as model-based and heuristic-based. Model-based methods construct a user profile from the existing data set where heuristic based methods use similarities among neighbors to construct a profile (Breese,1998; Adomavicius & Tuzhilin, 2005).

The advantages of using CF can be explained in the following way; since CF uses ratings of users, it finds correlated people exploiting user-item ratings matrices. Therefore, in CF there is no need to keep a record of the attributes of the item (Chen & Aickelin, 2004). The challenges of this kind of filtering are sparsity and scalability (Claypool et al., 1999; Sarwar et al., 2000). Cold-start is another problem of CF (Schein et al., 2002). It occurs when there is no information to make a decision about an item or user.

2.1.2. Content Based Recommender Systems

Content-based filtering (CBF) is another important technique used in recommender system research. CBF is simply the matching of user profiles and item attributes (Wen, 2008). In this type of recommendation, unlike the CF, the similarity of a given item and the other items preferred by the users in the past is investigated. The item which has the strong similarity with the given item is recommended as a result (Shih et al., 2011). Therefore, different from the CF, CBF gives importance to the item attributes rather than looking at similar users. The attributes of an item are the characteristics, which define it very well. For example, a movie can be best described by its attributes such as its title, its director, its genre, its actors and its native country. CBF takes into account the past item preferences of the users and the attributes of these items, because it assumes that these preferences would be the clues for the future activities of the users (Shih et al., 2011). Therefore, in CBF, there is always a need for information about the items. The advantages of using CBF is it provides good recommendations if the attributes of the items that would be recommended are being kept. A content analysis therefore needs to be done in these types of recommendation techniques. This particularity can easily become into a disadvantage if these attributes are not available. The reason is that, since this type of recommendation base it's filtering with respect to the content and therefore if the content is unavailable, the recommendation fails. Also, new user cold start problem which is present in CF is also a drawback for this type of recommender systems. Both content-based and collaborative filtering has the trouble of cold-start, sparsity and scalability (Ghazanfar & Prügel-Bennett, 2010).

2.1.3. Hybrid Recommender Systems

Hybrid recommender systems are known as the combination of two or more techniques. The idea behind this combination of the techniques is to generate the improved recommendations while exploiting from the advantages of the techniques, and overcoming the challenges of each technique with the aggregation (Burke, 2002; Ghazanfar & Prügel-Bennet, 2010). Burke (2002) presented the following hybridization methods to combine two or more techniques to build recommender systems: weighted, switching, mixed, feature combination, cascade, feature augmentation, and meta-level. Many of recommender systems recently studied are hybrid systems which combine collaborative and content-based filtering.

2.1.4. Demographic Recommender Systems

In this specific study, the emphasis is given to demographic recommendation technique which is enhanced with the user feedbacks by taking into account the user's past movie genre preferences. In demographic recommender systems, the recommendations are generated by taking into account, users' demographic characteristics and users' ratings (Krulwich, 1997; Pazzani, 1999). It is stated that demographic user segmentation is popular in marketing and in many websites (Ricci et al., 2011). Demographic Recommender Systems classify users with respect to their individual characteristic attributes and generate the recommendations based on demographic categories (Burke, 2002). Ricci et al. (2011) states that these kinds of recommendation techniques hypothesize that different recommendations should be given to different "demographic niches". Aimeur et al. (2006) claims that the most important feature of the Demographic Filtering is to group the users who have the same demographic characteristics into the same classes and to discover the buying behaviors or preferences of these classes. From this perspective, demographic filtering shows similarity with CF which also searches the similarity among the users but different from demographic filtering CF investigates similarity from the user ratings. The input for the demographic kind recommender systems is the user's demographic information; the task of the recommendation algorithm is to find users which are similar to the aforementioned user. An item can be recommended by looking at the similar users to the aforementioned user and how these users rated the item (Ricci et al., 2011). Recommendation generation for a new user is done by firstly putting this new user into the most suitable class by looking at his demographic characteristics, secondly assuming the same buying behavior or preferences of the users who were present before this user in that class and thirdly offering the same recommendations that are offered to that class, to the new user.

Krulwich (1997) and Pazzani (1999) used the users' data which contain the user demographic information in their systems. Demographic approach has some drawbacks, first it is difficult to collect personal data; second, since the recommendation is based on finding similar demographic profile, the recommendation may remain too general (Montaner et al., 2003); third, demographic attributes are stable, however people change over time therefore an adaptation is not possible (Koychev, 2000). Ricci et al. (2011) also draws attention to the fact that demographic data is sensitive to collect and they point that it is not convenient to say to people that they may be interested with some kind of movie, for example romance type movies since they are women. Another claim is that people's mental age and actual age may not be the same (Wang & Zhou, 2012). These challenges can be shown as the reasons why there are comparatively less studies in demographic filtering than other methods. Therefore, researchers are inclined to combine demographic data and other techniques more. Generally researchers has gone to the way to combine the demographic data with other data like user ratings, item features and so on to make recommendations especially to improve the recommendations (Ghazanfar & Prügel-Bennett, 2010; Vozalis & Margaritis, 2004).

Said et al. (2011) performed a comparison of how the demographic data affects the recommendation by using different demographic features. In their study, they assumed that demographic data stores implicit information about the users' interests, preferences, and choices. For example Chen & Le (2009) proposed a method in which a collaborative filtering approach based on demographic data is studied. This study is similar to our approach by calculating user similarities with the help of users' demographics rather than user ratings. The domain was movie as in our study. We perform a different study from the mentioned study, by not performing pair wise similarity calculation. We calculate similarity between a user and other users and perform this similarity calculation with the help of NBC.

2.2. Bayesian Classification and Recommender Systems in the movie domain

Billsus and Pazzani (1996) classified the websites based on the NBC. In their study it is shown that accuracy is not increasing with the enlargement of the database. Melville et al. (2002) studied the recommender systems in the movie domain and proposed a hybrid recommender system where they include the content information with the users' rating to recommend movies. They adopt this approach to avoid the drawbacks of CF such as sparsity and first-rater problem. To overcome the first-rater problem, content-based prediction of other users is used to find the similar users with the current users. They applied the NBC to learn a user profile from rated movies where this profile will be used later to predict the ratings of the movies that are not rated before. For this, they took movie features such as title, cast etc. With this study, they showed that this approach performs better than the pure collaborative filtering, content-based filtering and naïve hybrid approaches.

Markellou et al. (2005) also applied NBC to classify people according to their demographics combined with the movie ontology and reported that by observing users' demographics, behaviors, preferences, and ontology of the items and combining these with mining techniques, improves the recommendation success. Different from our approach they make use of movie ontology, and they did not apply any techniques for feature selection. Here, we use only the movie genre.

Moreover, they did not take into account the rankings of the users. Robles et al. (2003) points out that NBC is one of the most successful algorithms in many of the classification domains.

Ghazanfar & Prügel-Bennett (2010) proposed a recommender system where they combined the NBC and item-based collaborative filtering. The proposed system has the ability to switch between different recommendations techniques according to the prediction confidence. The idea behind this approach is to benefit from the strengths and minimize the weaknesses of the techniques. Different from the other studies and our study, researchers used Document Frequency Thresholding approach to select significant attributes for the feature selection method. Moreover, unlike our study, they have tags, actors, actresses, directors, plot, user comments, genre and synopsis data of a movie where they use NBC to predict the class of a movie.

CHAPTER 3

METHODOLOGY

In this study it is aimed to build a self learning recommender system by considering users' demographic attributes, day and time preferences and feedback. As a dataset, the real historical records of an online e-ticket entertainment company are used. This online company provides tickets from various entertainment activities ranging from theatres, cinemas, sports, events, concerts, shows and travels. In the dataset, there were historical records of users' movie genres choices. The grouping of people according to their demographics and assigning them to a movie genre category can be handled by the supervised learning technique which is a technique of Machine Learning Techniques (Kotsiantis, 2007). There are two learning types of Machine Learning techniques: Supervised and Unsupervised Learning. If the label for each instance is available in the dataset then the learning is called supervised whereas if the labels for these instances are not known, then the learning is said to be unsupervised (Han & Kamber, 2006). Classification is a type of a supervised learning, in which the outputs of the instances are discrete and unordered values (Kotsiantis, 2007). Amatriain et al. (2011) defines the classification as "a mapping between a feature space and a label space".

In the literature, many classification algorithms are adopted in the recommender systems research studies. The well-known classification techniques are Nearest Neighbors, Decision Trees, Rule-based Classifiers, Bayesian Classifiers, Neural Networks, and Support Vector Machines. The class labels are categorical values which are also called categories or classes. Each instance, *X*, can be represented by a vector with n dimensions. The dimensions are attributes and a class label exist, for each tuple *X* = *(x1, x2… xn)* (Han & Kamber, 2006).

In our study, each movie genre represents different classes. The classification task involves two main steps. The first step is to construct a model from the training set and the second step is to test the model by using the test set. For this reason the data is needed to be divided into two subsets such as training set and test sets (Han & Kamber, 2006). The working mechanism of the classification is given in the Figure 2 (Han & Kamber, 2006). Classification procedure is explained in following subsections in details.

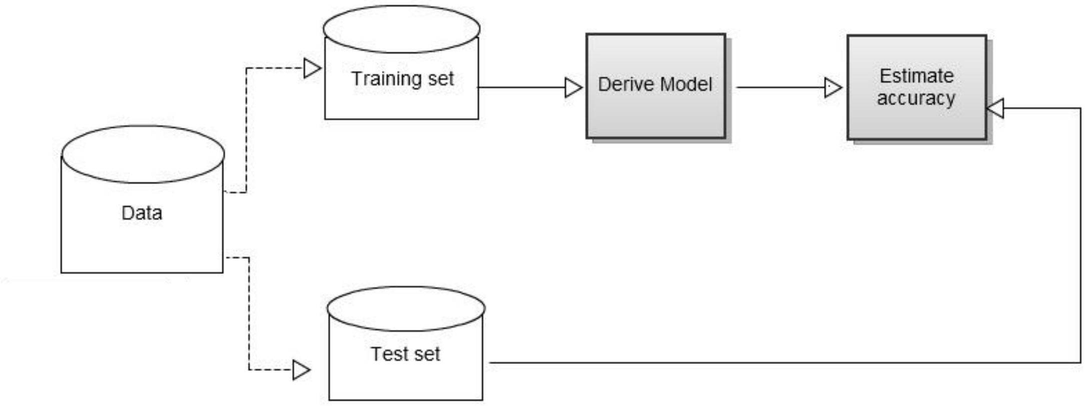


Figure 2: Classification working mechanism

Therefore, for our study, the supervised classification is found to be more appropriate. In this study, as a methodology, the supervised machine learning approach which is given by Kotsiantis (2007) is adapted to the present problem. The stages followed in the methodology are shown in Figure 3.

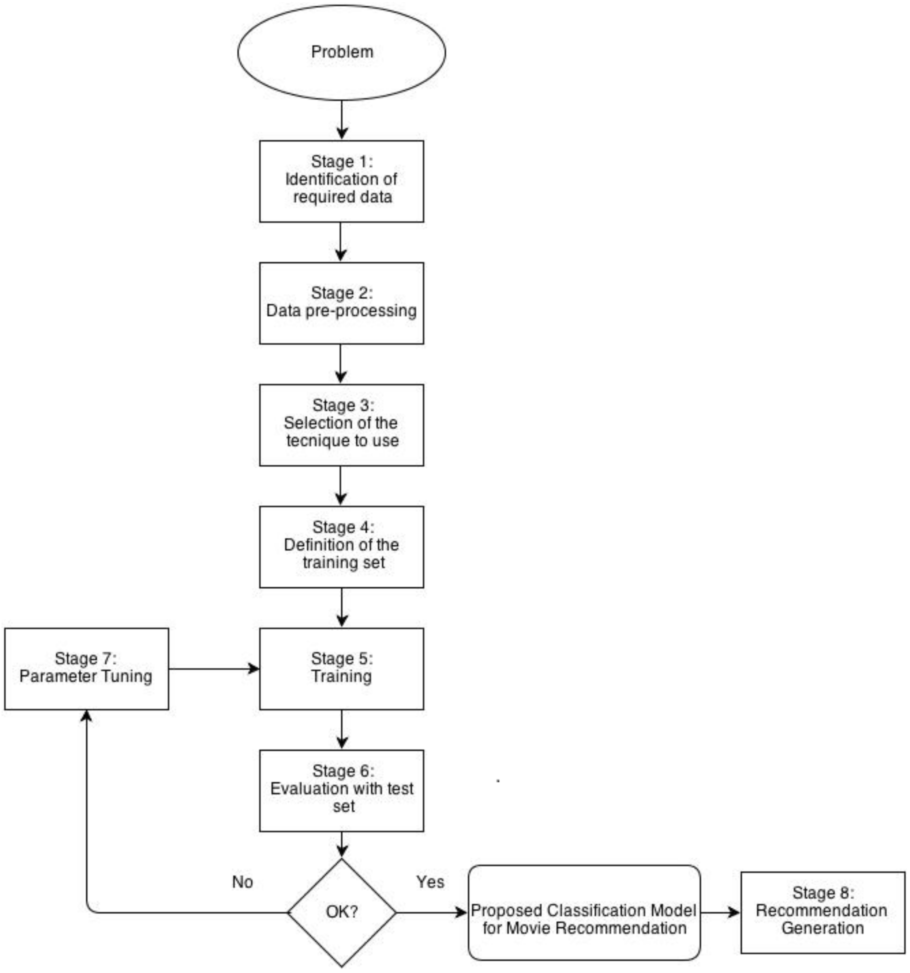


Figure 3: Stages followed in the methodology

3.1. Identification of required data

In this study we started from the idea that demographic information of the users such as gender, age, education, occupation, income, marital status, place they live etc. and also some external identifiers such as day of the week, time slot of the day and accompanying people affect the movie genre choice of people since different social groups have different movie choices (Said et al., 2010) and demographic, personality, social, and psychology based attributes affect media preferences of people (Kraaykamp & van Eijck, 2005; Nabi et al., 2006; Sargent et al., 1998).

Later, we diminished this set, by looking at the similar studies performed in the literature and giving also our arguments. It is stated that there are many web-sites providing recommendations by using demographic features of the users. As an example, the language they use and their ages are informative about their profile. In recent studies of the recommender systems' domain demographic information is used generally to construct a user profile in addition to the major techniques such as collaborative filtering and content-based filtering. (Ricci et al., 2011) pointed that collecting demographic information like age, gender, job area, nationality, language is an important step in the development of recommender user models because with the help of this attributes, the relationship among the users is provided. The demographic attributes taken into account in these studies are generally age, gender, education, occupation. Exemplar studies are as follows, Sobecki (2006) considered age, gender, and education, Ujjin & Bentley (2002) took age, gender and occupation attributes of the users by using a genetic algorithm in recommendation generation for the movie domain. Wei et al (2007) mention in the survey they have performed for the e-commerce recommender system, that demographic data of customers can be as follows; name, age, gender, profession, birth date, telephone, address, hobbies, salary, education experience and so on. Chen & He (2009) also utilized age, gender and occupation in his study where they used these demographic data with the collaborative filtering method to provide a solution to the cold start problem in recommender systems. Similarly, Chikhaoui et al. (2011) used age, gender, occupation and zip code in their recommender system study where they proposed to combine the collaborative filtering, content-based filtering and demographic filtering. Said et al. (2011) claim that the commonly available information in the current systems are the age, gender and location of the users, and concluded even a simple usage of this kind of data makes good improvements for the recommendation quality. Wang & Zhou (2012) have taken the age (7-17, 18-30, 31-45, and 46-60), gender and occupation (manager, artist, educator, engineer, and executive) in their recommender system approach where they used collaborative filtering. By looking at the presented studies, we therefore decided to choose age, gender, education, occupation attributes as the demographic attributes. In addition to the literature we have some reasons behind choosing these attributes:

Firstly, the demographic data is a private kind of data of users. The main reason that the researchers prefer to work comparatively less on demographic filtering than other filtering methods is the difficulty of collecting this private data from users. People generally are not willing to share their own data with others especially with the Internet sites. Also they are likely to be in a mood to give either false data (Montaner et al., 2003) or leave the spaces empty in the forms. Kotsiantis et al. (2006) points out that "Incomplete data is an unavoidable problem in dealing with most real world data sources." Therefore, due to the privacy concerns, it is not easy to collect much demographic data of people. Here, by considering the availability of the attributes, we eliminated some of them such as income, marital status, political view etc. As an example, it is not a frequent case that people share their own salary with websites because many of the users hesitate to share this kind of information with third parties.

However, with the diffusion of social media tools, blogs, and forums, some attributes of this private data change from being private to open-to-public. Attributes like age, education, occupation, gender become the most open and general attributes that can be gathered easily. Many of the main stream social websites such as Facebook, Twitter, LinkedIn, and so on collect at most up to this level of people attributes in their sign-up or later profile creation stages. Also, these types of social websites enable people to publicize their own profile over the Internet. If we give an example, from the graduation year of the university or high school it is not difficult to extract a user's age, and from the university they graduated or institution he/she is working, the occupation of the user can be easily gathered. Consequently, even though, earlier people were hesitating to share their private data, nowadays people are familiarized to share their gender, age, education, and occupation data which made these data to become open-to-public.

Secondly, apart from the availability of the demographic attributes, the effect of these attributes in this area of the study is also an important criterion for the elimination of some attributes. We eliminated first marital status, accompanying people, and place attributes due to the scope of the study. For marital status attribute, it is difficult to get the marital status of people, additionally it makes no sense in an internet based selling because a person may book a ticket for himself and for his wife but the transaction is recorded over one name and it is difficult to analyze this. Similar case is valid for the accompanying people attribute. Moreover, this study aims to predict movie genre choice of people but this can be applied in an online cinema ticket selling site or a DVD buying site. People prefer either to go to the cinema and they choose the movie theatres with high quality and near their homes, or watch the movies in their homes. For this reason, the place that a person lives in is likely to not being a distinctive mark for the choice of movie genre. Therefore in this study, we found the gender, age, education and occupation attributes as affecting factors of the movie genre selection.

The attributes of the movie itself were not included in this study. We could say that movie choice estimation of people can be affected from various attributes such as genre of the movie, director, country, actors and actresses, whether the movie was awarded or not; even the title of the movie is an important factor affecting people before watching a movie. In this study, since we are dealing with a demographic recommender system, the attributes of the item to be recommended is not a thing to consider. We only focused on the genre of the movie. The reason for this kind of choice is explained with two reasons in the following ways. First, we use a real historical user purchase data and in this data we do not have information of movie other than genre. Second, the genre of the movie is one of the most frequent descriptive attributes that people use to define the movies that they like or not prefer. Third, directors, actors and actresses or award information about the movies are some attributes that can be considered only by the people who have a high-level cultural knowledge/ interest about movie domain but people are likely to choose to go to the movies by looking either to the picture of the movie or by watching the trailer of the movies where the genre of the movie is informative about the movie. Therefore, these attributes are not likely to address entirely to the public. Among the movie genres, only six of them are taken into account and these are the main-stream movie genres: action, comedy, drama, horror, romance, and science-fiction. Although there are other genres as documentaries, war movies, film-noir etc., these types are not as frequent as the main-stream movie genres. Moreover, some genres such as animation, 3D are also excluded due to being the presentation type of the movie rather than genre.

In addition to the demographics, contextual attributes are taken into account in that study. People having different demographics may have different day and time preferences. For example, working people cannot prefer to go to cinema during the day times in the weekdays, similarly non-working people do not have this kind of restrictions. Basically, the key component of all recommendation systems is the user model which contains knowledge about the users' choices, preferences, and past activities which determine his behavior, in other words, his activities on the web.

The recommendation systems working mechanism can be summarized in two steps: user model construction and recommendation generation. We used the demographic attributes to construct a user profile by assuming that people with the same demographics will tend to behave in the similar way. Similar assumption is used in another study held by Chen & He (2009). There are two different ways of offering recommendations to the user: first one is to predict the ratings of the items that the user did not see before, second one is to compose an ordered list with respect to the preferences of the user (Ghazanfar & Prügel-Bennett, 2010). In this study, we firstly construct a user profile with the help of demographic attributes and then add the feedback of the users to make the system that learns from the previous interactions. Therefore, we concentrated on the second way by taking into account the movie genre rankings of the users and focused to the problem of grouping the demographically similar users.

3.2. Data pre-processing

Once the data that will be used is found, the next step is the pre-processing of this data. Data pre-processing stage forms an important part of the total work. Since both model construction in other words training and later testing depend on the data, the preparation of the data such as data cleaning, reduction, in other words, removing the incomplete, noisy and redundant records is required. It is reported by Zhang et al. (2003) that the data cleaning and preparation activities constitute 80% of the total data engineering effort eliminating less important attributes ameliorates the computational time (Ting et al., 2011). Especially, in real circumstances, generally people are not likely to fill the forms, share their private data with others. For this reason they either do not provide some data or provide wrong data, for example they do not enter their real age and just click one option from a combo-box. Han & Kamber (2006) points out that "Low-quality data will lead to low-quality mining results". For data mining, the quality of the data is more important than the size of the data. High-quality data is the one which does not contain noisy and missing records, which can be grouped and which are scalable (Oz, 2007). Therefore, a significant effort is given to pre-process the data. It should be emphasized that cleaning the data is the core step in data preprocessing. Since our dataset is composed of the real purchase history records of an entertainment company, a cleaning need is expectable. Moreover, it consists of the activities other than cinema; all the records related to other activities are removed from the dataset. After that, the steps for cleaning the data are extracted and adapted to our case from the approaches given below that are suggested by Han & Kamber (2006) and these steps are followed to pre-process the data. The steps followed in this study are as follows:

Step 1: Handling the Missing Values: The original data was composed of many missing values. To handle the missing values it is given 6 different approaches such as "ignoring the tuple", "filling the missing value manually", "using a global constant to fill in the missing value", using the attribute mean to fill etc. Among the given approaches for handling the missing values ignoring the tuple approach is adopted for getting accurate results, it is decided to take only the available data. Hence, the tuples in which a missing value exits were ignored.

Step 2: Smoothing noisy data: To smooth the noisy data, some questions provided by Han & Kamber (2006) were answered. Some of the questions are "What are domain and data type of each attribute?", "What are acceptable values for each attribute?", "What is the range of the length of values?" etc. Some values were containing noises such as -1, 0 for texts. These records are considered as outliers and are excluded from the dataset.

Step 3: Resolving inconsistencies: In this step, a look for inconsistent data representation is done and many inconsistencies are found to be resolved. For example for the dates representations such as ".", "/", "-", were available in the dataset and all these representations were replaced by ".", Similarly, for the districts, there were named in various ways, for example, "Çankaya", "Cankaya", "Çankara" all of which is "Çankaya", because of wrong entrance of the user during their registering, all these different naming are corrected. Moreover, to use the data later in the analysis, two words separated by a blank are combined and also the Turkish characters from the words are replaced with the equivalent characters from the English alphabet such as "Ü" which is replaced by "U". Last but not least, even though it seems that the age attribute is acceptable as well as the education attribute, there were some records in which the combination of both was not relevant; as an example, a person whose age is 19 has an education entry such as PhD. Since the number of this kind of records were not too high, these kinds of tuples were also eliminated.

Step 4: Elimination of the less significant attributes: Since the dataset is composed of the historical records of an electronic ticket company of entertainment, in the data there were some contact information of the users such as e-mail, phone and some transaction information like the server-id, whether the user used a discount, the type of the discount used. These attributes were removed from the dataset. For the place information, the data had only the district information. Since the districts are too various, they cause a high-dimensionality for the data. The only thing that may be done was to categorize them according to the city or region they belong to. However, as it is anticipated and explained in previous section, it will not give a significant idea about the movie choice. People generally; prefer to watch the movies in the cinemas that are near to them or in their homes. For this reason, place attributes were eliminated too.

Step 5: Data Integration: The data exist in four sheets and for this reason the data was integrated in a single sheet. The data was merged by using Excel Vlookup function from different tables by examining the unique rule.

Step 6: Data Transformation: In the data transformation step, as an example, there was age attribute in the dataset and the ages are grouped into 3 categories from smaller than 30, between 30 and 45 and higher than 45. The hour attribute also is transformed to a categorical type by giving categories to the different time slots of the day. Same procedure is applied for the days of the week and the days are categorized into 2 categories such as weekend and weekday. Also, there were 51 different occupation categories available in the dataset which were decreased in groups of 9 by creating general categories. These categories are finance sector, service-commerce sector, health sector, technology sector, managers, students, instructors, and others.

Step 7: Identifying / removing outliers: There were only 35 records for documentary, 22 records for war genres which were removed from the dataset since their occurrence will not likely to be observed due to the limited study time. Therefore availability of these types of records will be insignificant.

3.3. Selection of the technique to use

In this section, the adopted technique: Naïve Bayes Classifier is explained along with the reasons why it is adopted.

3.3.1. Reasons for adopting NBC technique

Bayesian Classification is called "Naive" since in this type of classification; it is assumed that all the attributes which are available in the problem domain are independent of each other (Dunham, 2003). Although the Bayesian classifier is an old technique that was studied by many of the researchers, it is still preferred by the researchers recently. For example Engelbert et al. (2012) adopt the Bayesian Classifier for the TV Recommendation system. Similarly, De Pessemier et al. (2010) extend the Bayesian Classifier to construct a context-aware recommender system for the mobile devices. Ghazanfar and Prügel-Bennett (2010) combined the Naïve Bayes classification technique with the collaborative filtering. Ting et al. (2011) in their study investigated whether the NBC is a good classifier for document classification. They reported that NBC outperforms other classifiers such as decision tree, neural network and support vector machines in document classification by comparing the accuracy and computational efficiency. Other reasons that directed us to choose the NBC as follows:

Firstly, since in this study we deal with the categorical data, it is appropriate to use NBC which is a classification technique that is suitable for categorical data types. Additionally, it is stated by Sharma & Mukherjee (2012) that NBC technique is appropriate for the cases in which the dimensionality is high. The dimensionality is the number of input attributes. In this study the number of attributes is six. These are demographic attributes: age, gender, education, occupation as well as contextual attributes: day of the week and time slot of the day. Each attribute has some sub categories. The attribute details are given in the Table 1. Secondly, NBC assumes that all the attributes are independent of each other and have equal impact on the classification problem (Ting et al., 2011). NBC makes the classification by combining the effects of the attributes. Therefore it can handle missing attributes during the classification, since this method omits these missing values while performing the calculation of the likelihood (Dunham, 2003). NBC has a capability of ignoring the missing attribute due to its conditional independence assumption. Since the developed approach can be adapted to the market, the missing attributes are inevitable and should be taken into consideration.

On the web, many websites ask people to fill many fields in the forms especially in the sign up stage. It is quite common that people do not fill all the blanks in the forms. In other words, users on the web usually tend to skip the areas that are supposed to be filled, (Montaner et al., 2003). Kotsiantis (2007) reported a comparison of different classifiers such as NBC, Decision Trees, Neural Networks, k- Nearest Neighboring, Support Vector Machines and among these techniques the classifier which has the ability to handle missing attributes much more than the others is found as NBC. Robles et al. (2003) pointed out that this particularity of NBC provides researchers to use only the features whose values are available for the classification. This characteristic of NBC makes it powerful in the conditions where some of the attributes are missing.

Table 1: Chosen attributes and their values

ATTRIBUTES and VALUES	
Day	Weekday
	Weekend
Hour	11:00-15:00
	16:00-17:00
	18:00-19:00
	20:00-24:00
Occupation	Finance
	Service-commerce
	Student
	Instructor
	Health
	Technology
	Manager
	Other
Education	High-school
	University
	Graduate (m.sc. or Ph.D.)
Gender	Male
	Female
Age	Below 30
	Between 30 and 45
	Above 45

Thirdly, it is stated that with the usage of NBC, it is easy to observe the learning and single scan of the training data is sufficient, where this is not the case in other methods. Moreover, Berry & Browne (2006) points out another important advantage of using NBC is that it is easy to verify, and if the data set is large enough, the classification will be performed in a consistent way and with minimum error.

Fourthly, when the literature is investigated, it is seen that although there are many recommender system applications that adopted NBC technique, it is discovered that none of them used the DT enhanced NBC for the feature selection. Moreover, even though the DT enhanced NBC is tested and it outperforms the traditional NBC in many domains, it is not tested in the movie domain which motivated us to study this technique in this thesis. The working mechanism of Naïve Bayesian Classifiers is given in section 3.4.3.

3.3.2. Reasons for not adopting other classification techniques

Among different classification techniques, we preferred to use Naïve Bayesian Classifier (NBC) due to its success despite its simplicity (Russell & Norvig, 2002; Sharma & Mukherjee, 2012). We did not prefer to use:

- Decision Trees and Rule-based approaches as classifiers; since they are not very tolerant to redundant and irrelevant attributes; they are not as successful as NBC in incremental learning (Kotsiantis, 2007). It is difficult to base recommendation model on rules (Amatriain et al., 2011). Moreover, decision trees work in divide-and-conquer strategy and therefore are not efficient (Elomaa & Rousu, 1999). Also, it is difficult to implement a learning decision tree, since they do not support the online learning and moreover it is stated that building an optimal decision tree is NP-complete (Tan et al., 2005). It is also stated in Amatriain et al. (2011) that it is unpractical and hard to implement a DT and it is better to use DTs as the parts of the systems.
- Bayesian Networks as classifier; actually NBC is a type of a Bayesian Network in which all the attributes are considered independent and no parent-child relationship exists among the attributes (Jiang et al., 2009). However, building a Bayesian Network is NP-hard class problem; for this

reason people tend to use NBC instead (Chickering, 1996) and since it is aimed to implement a prototype recommender system, it will not be a good choice to base the system on a Bayesian Network classifier.

- Neural Networks as classifier; they are not tolerant for missing values, tend to over fit to the data, and model parameter handling is low (Kotsiantis, 2007).
- K-Nearest neighboring as classifier; they are not tolerant to missing values and redundant attributes and classifier is sensitive to the choice of the value of *K* (Dunham, 2003).

3.3.3. Naïve Bayesian Classifier

Naïve Bayesian Classifiers are statistical classifiers that can guess the probability of a given tuple belonging to a specific class (Han & Kamber, 2006). Naïve Bayesian Classification is based on Bayesian reasoning. The elements of Bayesian approach for the classification are conditional probability, Bayes Theorem and Bayesian Decision Rule. Bayesian reasoning is used in the domains where uncertainties exist. Especially, in real-life, uncertainty is inevitable. Attributes' conditional independence is the assumption that NBC relies on; this assumption can be explained in the following way that the impact of an attribute, on a given class is independent of the impacts of other attributes. Bayes' Theorem is given in Equation 1, where *P (Y|X)* denotes the posterior probability for the class *Y*; *P (X|Y)* denotes the class-attribute conditional probability; *P (Y)* denotes the prior probability of *Y*; and *P(X)* denotes the evidence record being classified.

$$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)}$$ **(**Equation 1)

The posterior probabilities for each class *Y*, and for attributes $X=\{ x_1, x_2,..X_d \}$ can be calculated as follows in Equation 2:

$$P(Y|X)=\frac{\prod_{i=1}^{d} P(X_i|Y) \times P(Y)}{P(X)}$$ (Equation 2)

The value in the denominator of the formula, *P(X)*, always remains constant. Consequently, it can be omitted during the calculations (Tan et al., 2005) and the formula evolves to the following Equation 3:

$$P(Y|X)=\prod_{i=1}^{d} P(X_i|Y) \times P(Y)$$ (Equation 3)

When using the NBC, the aim is to find the class having the maximum posterior probability among each different movie genre classes, and to assign the given record to the class having the highest posterior probability.

Naïve Bayesian Classifier also provides a robust solution to the zero probability problems. Zero probability problems occur when the conditional probability of an item given a class is zero, Laplace correction or m-estimates can be used to alleviate this problem (Dunham, 2003).

3.4. Definition of the training set

The most important thing in defining the sample is, choosing the sample which is not bias and as representative as possible. Both for the training and testing, it is needed to take a sample from the whole data. Sampling needed to be done since using the whole dataset is computationally expensive (Amatriain et al., 2011). Therefore, sampling is an important step of data mining applications, and the chosen sample should reflect the original big data set as much as possible. As sampling method, stratified random sampling is used to choose the study data from the entire set of data. Entire pre-processed data set consist of 6000 records but to make every movie genre classes to appear in the sample homogeneously 1446 records are sampled. A similar document classification study was held by Ting et al. (2011) and they classified documents into four sets using NBC where they used equal number of instances for each document category. As stated in the previous section only six genres which are main-stream genres are taken into account in this study. The sampling is performed to take as much as possible number of records but since equal number of movie genres are decided to be taken, the number of each class is chosen according to the smallest number of record for a genre which was 241. Hence 241 x 6= 1446 records are sampled. Appendix E shows the demographic distributions of the sample set.

To identify the testing and training sets, stratified random sub sampling is used; the number of records for each category remained the same i.e. homogeneous in the sets. Random sub-sampling is *k* times application of the holdout method; where in the holdout method, the data is divided into two subsets, which indicate the training and testing sets. Random sub sampling is preferred because, sampling may cause over-specialization to the particular split of training and testing data and therefore it is better to repeat it several times (Amatriain et al., 2011). Researchers generally suggest to use any value over 2/3 for the training set and the remaining third for the testing purpose (Han & Kamber, 2006; Kotsiantis, 2007; Amatriain et al., 2011). Therefore, we chose to apply the random sub sampling method 3 times by choosing 2/3 for the training set and the remaining third for the testing purpose.

Another method for the evaluation of the classifiers is the cross-validation. In this method, the data is divided into k subsets which are usually called “folds” and training and testing is done k times. This method is not used in this study since for the evaluation we use our system, not a readymade tool like Weka or others. For this reason, testing the folds is not efficient if the fold numbers are large. Moreover, it is stated in (Amatriain et al., 2011) that cross-validation may not be reliable if the data set is not large and random sub sampling is generally acceptable.

3.5. Training the data

With the help of the conditional independence assumption of NBC, the conditional probability of each attribute $X_{i,}$ given Y is calculated. NBC is based on the probabilistic calculation of the effects of each attribute to the target variable. In this study, all the people in the training data are categorized into six different categories in which the model can simply compute which person is likely to prefer which movie genre category. In this study, as stated before, the classes are the movie genre which are *Y= {drama, comedy, horror, romance, fiction, action}* and the attributes are *X= {age, gender, occupation, education, day, time}*. To calculate the conditional probabilities, contingency tables are used. Contingency tables are useful tools showing the relations of two variables that can be contingent, in other words, dependent on one another. The contingency table is provided in Appendix F. We decide whether a person chooses to go to the drama movie according to the impact of each attribute of this person such as his age, education, gender, occupation, day and time preferences. Consequently, when we give as input a person with specific attributes to our system which is trained by the probabilities calculated before, the system will predict whether this person will likely to go to a drama, romance or other movie types. In other words, the probability of occurrence of a result is the product of all the attributes which provides this probability.

Suppose a test record *X = { gender=”female”, education=“high school”, age= “below30”, occupation= “student”, day= “weekend”, hour= “18:00-19:00”}*. To classify X to the movie genres, we have to calculate the posterior probabilities:

P (fiction | X), P (drama | X), P (horror | X), P (comedy | X), P (action | X), and P (romance | X). After calculating these probabilities, the test record *X* will be assigned to the class of movie which is the maximum among these posterior probabilities. Then, the output of the NBC is *P (Y | X)* the probability of a person *X* belonging to a movie genre class *Y*. According to the equation 2, the posterior probabilities are computed as follows:

$$\begin{aligned} P(comedy|X) = P(female|comedy) \times P(highschool|comedy) \times P(below30|comedy) \\ \times P(student|comedy) \times P(weekend|comedy) \times P(evening|comedy) \end{aligned}$$

For each movie genre, the same calculation is followed and we finally compare the probability results. After the comparison, the record *X* is assigned to the movie genre class which has the highest probability among others.

In this stage, we make an assumption that a movie genre preference is not a unique choice. It is not true to reach an argument that people who love drama do not prefer horror movies. They can prefer both. In this specific study it is suggested to people not only the movie genre which has the maximum probability, but also the genres in the second level. Therefore, among six different movie genre categories, matching two of them are checked as classification results. In the literature, this fact is called as “second guess” heuristic (Han & Kamber, 2006) and is used for the situations in which an instance can belong to more than a single class. Han & Kamber (2006) states more formally that “A class prediction is judged as correct if it agrees with the first or second most probable class.”

The evaluation of the model is performed in the following way: a tuple whose class label is known from a testing set is compared with the resulting class label produced for this tuple from the model which is constructed using the training set. The validity of the model is determined by the ratio of the truly classified examples over the total test set (Berry & Browne, 2006).

The aim of this study is to compare the classification approaches NBC and DT enhanced NBC in the movie domain and development and assessment of a recommender system according to the successful methods between the two methods. For this reason first of all, the classification accuracy of these two approaches needed to be investigated.

3.6. Evaluation with the test set

Recall from section 3.3 that researchers generally suggest to use any value over 2/3 for the training set and the remaining third for the testing purpose (Han & Kamber, 2006; Kotsiantis, 2007; Amatriain et al., 2011) and suggest to repeat the training several times to avoid over-specialization (Amatriain et al., 2011). When examplar studies are observed, it is seen that researchers divide data from 60 %-40% to 90%-10 %. Therefore the trainings and testing were done with changing size such as 60% for training and 40% for testing, 70% for training and 30% for testing, 80% for training and 20% for testing and lastly 90% for training and 10% for testing.

Accuracy measure is used to evaluate the success performance of the NBC (Tan et al., 2005). Berry & Browne (2006) pointed out that the measurement of the accuracy for the classifier can be found by looking at the proportion of samples that were classified correctly. Accuracy is the measurement of counting correctly classified class labels of an unseen data and is defined as:

$$\text{Accuracy}=\frac{\text{Number of correct predictions}}{\text{total number of predictions}}$$

To perform the evaluation, a java application is developed. The Naïve Bayesian Classifier is implemented and the conditional probabilities that stem from the training set are kept in the database. In the evaluation stage, each record of the testing set is given to the system input one by one. Movie genre label of each test record is compared with the movie genre class prediction of the classifier. A record's classification performance is always dependent on prior records, since when the actual movie genre and the predicted movie genre do not match, the probabilities are updated according to the actual class. With this way, a learning ability is achieved. The following Figure 4 adapted from (Han & Kamber, 2006) illustrates the process:

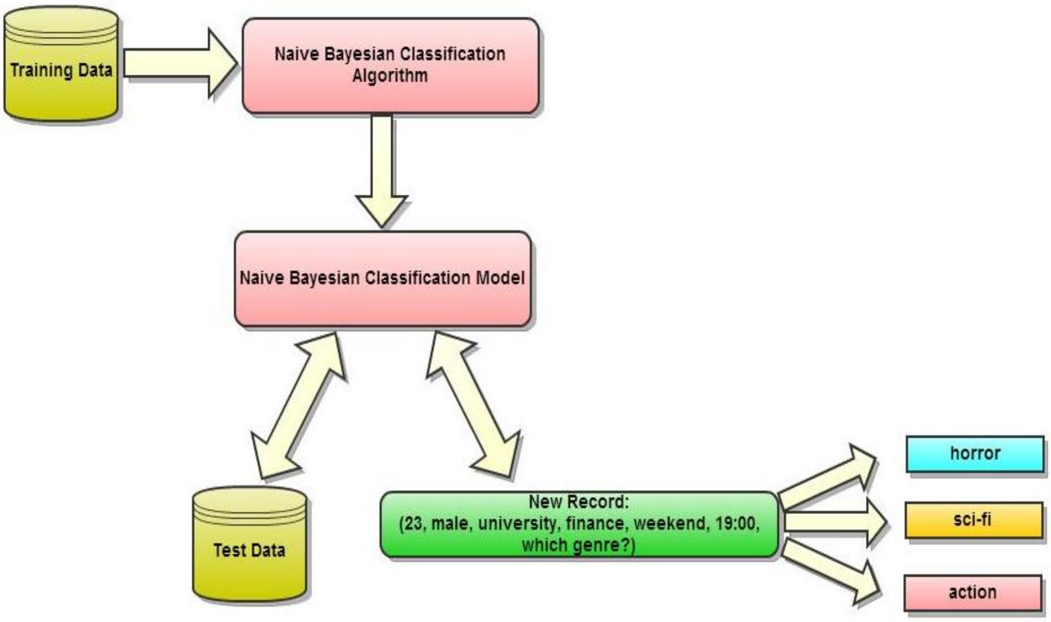

Training Data
Naive Bayesian Classification Algorithm
Naive Bayesian Classification Model
Test Data
New Record:
(23, male, university, finance, weekend, 19:00, which genre?)
horror
sci-fi
action

Figure 4: The classification process

The results of this evaluation are given in Table 2.

Table 2: Accuracy of NBC

Sample size	Technique
60% for training and 40 % for testing	**NBC**
First Genre	26 %
First & Second Genre	47 %
70 % for training and 30 % for testing	**NBC**
First Genre	19 %

Table 2 continued

First & Second Genre	36 %
90 % for training and 10 % for testing	**NBC**
First Genre	16 %
First & Second Genre	35 %

According to the results presented in Table 2, the accuracy of the NBC is not encouraging. One reason can be the nature of the dataset used. There is nothing more to do to the dataset in this stage. Another reason can be inherent from the nature of NBC. From the conditional independence assumption of NBC, each attribute are considered to be independent of each other. Although this assumption is accepted as one of the most important characteristic of NBC, it easily becomes a disadvantage when the attributes are redundant. To check this issue, we followed the alternative stage of parameter tuning.

3.7. Parameter tuning

The purpose of the study is to classify people by considering their demographics as well as contextual attributes such as day and time to an appropriate movie genre category. The adopted technique for the purpose is NBC. After testing the accuracy of the classifier, we found that the classification accuracy is not encouraging which can be seen in Table 2. Therefore a need for the improvement of NBC is arisen. As stated in previous section 3.3 and 3.6. NBC assumes that each attribute is independent of each other and each of them has an equal contribution to the classification activity. However, this conditional independence assumption of NBC also returns itself a disadvantage, since many attributes are dependent in the reality. Ratanamahatana & Gunopulos (2003) points out that NBC suffers from the redundant and/or irrelevant attributes because they receive duplicate probability effect which directs the model to make a wrong classification. Hall (2007) supports this idea that even a single redundant attribute which is correlated with another attribute has twice more influence than other attributes. Therefore, the basic idea is to eliminate the redundant and / or less significant attributes in the data set. To accomplish this aim, we used another well known supervised classifier Decision Trees (DT) to extract the most significant attributes. To put in another way, feature selection is required to eliminate the redundant attributes that the NBC uses and for this reason DT is used.

3.7.1. Feature Selection for Naïve Bayesian Classifier

Feature selection methods are used to eliminate the redundant, less significant attributes and use only the remaining attributes which have more significant power in the classification process. The available feature selection methods generally use DF-Thresholding, chi-square statistics, and information gain (Ghazanfar & Prügel-Bennett, 2010). To accomplish this aim, the literature is investigated and DT technique is found to be a successful technique whose success is proven in other domains (Ratanamahatana et al., 2003; Hall, 2007; Gayatri et al. (2010). Decision Trees are applied as a feature selection method for classifiers especially for Naïve Bayesian Classifiers in various studies.

First, Kubat et al. (1993) used decision trees in feature selection for NBC, to discover EEG (Electroencephalographical)-signals patterns. One derivation of combination of NBC and DT suggested by Kohavi (1996), is simply applying the NBC to the leaves of a DT. This approach is called NBTree and outperformed both NBC and DT classifiers (Kohavi, 1996). Later Ratanamahatana et al. (2003) proposed to use the decision tree algorithm to extract the significant attribute for the Naïve Bayesian Classification. Different from the study proposed by Kubat et al. (1993) in which just one tree is created from the training set and all the attributes appearing in that tree were considered as significant, Ratanamahatana et al. (2003) constructed many trees and suggested to take the attributes appearing in the first three levels of the decision trees in the NBC classification which they call Selective Bayesian Classifier. Their aim is to improve the performance of the NBC by eliminating less significant and redundant attributes. As a DT algorithm, C4.5 algorithm which is based on information gain was used. In this algorithm, attributes with higher information gain appears in the upper levels of the tree which are near to the root node. They tested their model in 10 wide ranges of datasets ranging from proteins, soybeans, to vote data provided by UCI repository[1]. They concluded that the Selective Bayesian Classifier is successful and outperforms the traditional NBC.

[1] http://archive.ics.uci.edu/ml/

Later, Hall (2007) offered another approach to improve the performance of NBC, similarly by using the feature selection. In this approach, Hall (2007) proposed a technique which combines the DT and NBC by giving weights to the attributes that are available in the DT and then using the NBC. In that study, the researcher sets the weights of the attributes by looking to the depths of the attributes in the tree. The weights were inversely proportional with weights. Naturally the attributes which do not appear in the tree receive zero weight. They made stable the estimated weights by constructing multiple trees and taking the averages of the weights. Like the previous studies by Kohavi (1996) and by Ratanamahatana et al. (2003), the model is tested on the datasets in UCI repository. They showed that this method outperformed the standard NBC.

However, to our knowledge, in the literature, there is no study which combines NBC and DT as hybrid approach in a demographic recommender system for the movie domain to recommend movie genres by considering users' demographic particularities and day and time preferences. Therefore, we find noteworthy to investigate whether this combined approach is successful in the movie domain as well as for a recommender approach. For DT algorithm we used two algorithms, CHAID and CART explained in details in the following sections.

3.7.2. Chi-squared automatic interaction detection (CHAID) Decision Tree

Decision trees can be implemented by various algorithms, some of them are well known in the literature as ID3, C4.5, C5.0, CART, CHAID, and QUEST. Among these well-known algorithms, we first chose CHAID (Chi squared automatic interaction detection) for the feature selection of Naïve Bayesian Classifier. The reason behind this choice is that, it is an effective segmentation technique, suitable for the cases in which the attributes are demographic and when the target dependent variable is categorical (Magidson & Vermunt, 2005; Dunham, 2003). This is important because for these kind of data, well known methods such as regression is not suitable and CHAID is therefore very useful to measure the relationship between attributes. CHAID uses the chi-squared statistics to find the splitting variable by using chi-square test. Best split is found by combining the pairs of independent variables if there is no statistically significant difference between them (Dunham, 2003). Chi-square statistic measures the dependency between two variables, determines significant attributes and merges the ones which do not differ while anticipating the target variable (Magidson & Vermunt, 2005). When the split node is determined, the same procedure is repeated for each remaining child nodes until there is no node that can be split, i.e. until no significant pairs are found. CHAID does not perform pruning (Bijak & Thomes, 2012). Despite other DT algorithms such as QUEST, CART which grows as binary trees; CHAID creates multiple branching trees. However, it is stated that it does not promise the best split at each node (Dunham, 2003). Moreover, it is stated that it is likely to be difficult to analyze the results inherent from the tree if the branching is high (SPSS, 1999). Decision Trees used in this study are constructed with help of the IBM SPSS Statistic Software Version 20 tool.

We used CHAID in two different ways; first we took the attributes appearing in the tree as significant attributes and applied the NBC by using these attributes which is given Table 3. Second, we followed the paths of the tree and applied NBC using the attributes appearing in these paths adaptively. This second procedure is also known as NBTree in the literature. Comparative accuracy results according to the CHAID enhanced NBC is given in the following Table 3.

Table 3: Accuracy results of NBC, CHAID & NBC and NBTree with different sample sizes.

Accuracy results			
Sample sizes for training and testing	**NBC**	**CHAID Decision Tree enhanced Naïve Bayesian Classifier**	**NBTree**
60 % for training 40 % for testing	**NBC**	**NBC+ CHAID DT**	**NBTree**
First genre	26 %	70 %	70 %
First or Second genre	47 %	83 %	83 %
70 % for training 30 % for testing	**NBC**	**NBC+ CHAID DT**	**NBTree**
First genre	19 %	25 %	27 %
First or Second genre	36 %	42 %	44 %
90 % for training 10 % for testing	**NBC**	**NBC+ CHAID DT**	**NBTree**
First genre	16 %	18 %	18 %
First or Second genre	35 %	41 %	41 %

From Table 3, it can be seen that the results are improved when DT is used as a feature selection method for the NBC technique in the movie domain like other domains that were investigated by Ratanamahatana et al. (2003) and Hall (2007). It should be noted that in the experiment conducted with 60% training and 40% testing data a huge increase occurred. The reason of this drastic increase can be the only one attribute appearance in the decision tree. Since one attribute appeared in the tree, therefore NBC made classification by considering only one attribute.

It can also be seen in Table 3 that the number of correctly classified instances decreases, when the number of samples for testing decreases. The reason of this fact can be the different distributions of training and test set. Moreover, it observed that there is no proper trend such as meaningful and encouraging changes when the NBC is applied to the rules extracted from the CHAID algorithm, in other words, when it used as NBTree.

Even though the accuracy results are improved by using CHAID decision tree for the feature selection of NBC, we do not prefer to use this CHAID algorithm for the feature selection of NBC. According to CHAID analysis, attributes that appeared in DT are found as age, hour and day. Since the objective of this study is to offer recommendation to users by the help of the demographic attributes, the CHAID approach is found to be not adequate for this aim. It brings only one attribute to be the significant attribute which is the age attribute, among four attributes age, gender, education and occupation. Hence, age attribute becomes the only demographic attribute to consider. However, our aim is to recommend people by considering their demographics and with this approach we move away from the focus. Because, the thing we want to do with CHAID is to extract the dependencies among variables by this way to eliminate the redundant attributes which shows the same effect for the overall result. When we look at the literature, researchers however showed that other demographic attributes may have impact on the movie genre choices of people or researchers took these demographics in their research and the studies that taken these attributes in their studies is already provided in section 3.1. Therefore, this fact directed us to adopt and investigate another DT algorithm called CART (Classification and Regression Trees) and CART is also known as a feature selection method for the Naïve Bayesian Classifier.

3.7.2. Classification and Regression Tree (CART)

CART is another DT algorithm which is developed by Breiman et al. in 1984. It is a binary recursive decision tree that is used to select the most important independent variables in determining the target variable (Questier et al., 2004). CART assumes that the training set consists of impure set of examples of different classes, and therefore tries to separate the set into more pure, as homogeneous as possible subsets. This procedure continues until the homogeneity state is achieved. While building a decision tree, the aim is to find an attribute that partitions the impure set into a number of partitions which are less impure than the original set. The attribute whose values decrease the impurity the most is identified as the root node of the DT (Du, 2010). Gini index is used to measure the level of impurity of the data set in which the response value is categorical. Gini index does not use probabilistic assumptions where ID3 and C 4.5 algorithm looks (Lavanya & UshaRani, 2011) which is the reason why in our study CART is selected for the attribute selection of NBC where we already use the conditional probabilities for the classification. The Gini index of a node with *n* objects and *c* classes is defined in (Questier et al., 2004) in the following way:

$$Gini=1-\prod_{j=1}^{c}\left(\frac{n_j}{n}\right)^2 \qquad \text{(Equation 4)}$$

Where n_j denotes the number of *n* objects from class *j* available in the node.

Unlike CHAID in which first tree level is important, CART can get important information in the lower levels. (Questier et al., 2004) points that disrupting the tree growing in some stage may result to uncover the interactions between independent variables therefore, it is suggested to let the tree grow to its maximum size and then pruning it to avoid overfitting. Overfitting is a problem arising when the model does not fit future states. Tree Pruning is used to discard the sub trees which may over fit the data, and improved the accuracy of the classification tree (Lavanya & Usha Rani, 2011). Another solution to avoid the over fitting in classification is to split the data set into training and testing sets which is already applied in this study. In the literature, there are studies that used CART as a feature selection method for NBC.

For example, Gayatri et al. (2010) used decision tree induction to select the relevant features for the classification of software defect prediction, CART was one of the decision tree algorithms they tested and compared with other feature selection approaches such as SVM (Support Vector Machines) and RELIEF feature selection techniques. It is found that DT induction based feature selection performed better than other traditional classifiers including NBC.

However, the studies did not address either the movie domain or a recommender system problem. With this study we also test whether this approach is convenient for the recommendation purpose in the movie domain.

When the same procedure with CHAID and NBC is applied to CART and NBC, the results in Table 4 are obtained. In this analysis, we did not test the accuracy for NBTREE approach because due to the nature of the CART, an attribute can appear more than one time in a single path. Since the tree searches for the homogeneity, an attribute can be further divided into two at the lower levels and sometimes a path can be made up of just one type of attributes. For this reason, we did not prefer to investigate rules in that tree. The accuracy measures of two different classifiers the CART and NBC are given in the following Table 4.

The improvements are similar to the previous studies in other domains by Ratanamahatana & Gunopulos (2003). However, the accuracy results still are not encouraging which can be explained with the nature of the dataset used. The reasoning about the dataset is given at Chapter 5 section 5.1. This model - NBC is tuned to select the features which are more significant using CART- eliminated only the gender attribute from the demographic attributes. After making experiments in 3 different sample sizes such as 70-30, and 80-20, and 90-10, frequency of attributes appearing in the CART are presented in Figure 5.

Table 4: Comparison of NBC and CART enhanced NBC in terms of accuracy over different sample sizes

Accuracy results		
Sample	**NBC**	**CART Decision Tree enhanced** **Naïve Bayesian Classifier**
70% Training **30% Testing**	**NBC**	**NBC + CART DT**
First genre	18.75 %	20.83 %
First or Second genre	35.87 %	37.03 %
80% Training **20% Testing**	**NBC**	**NBC + CART DR**
First genre	25.60 %	26.64 %
First or Second genre	43.94 %	46.36 %
90% Training **10% Testing**	**NBC**	**NBC + CART DT**
First genre	15.97 %	21.52 %
First or Second genre	35.41 %	40.27 %

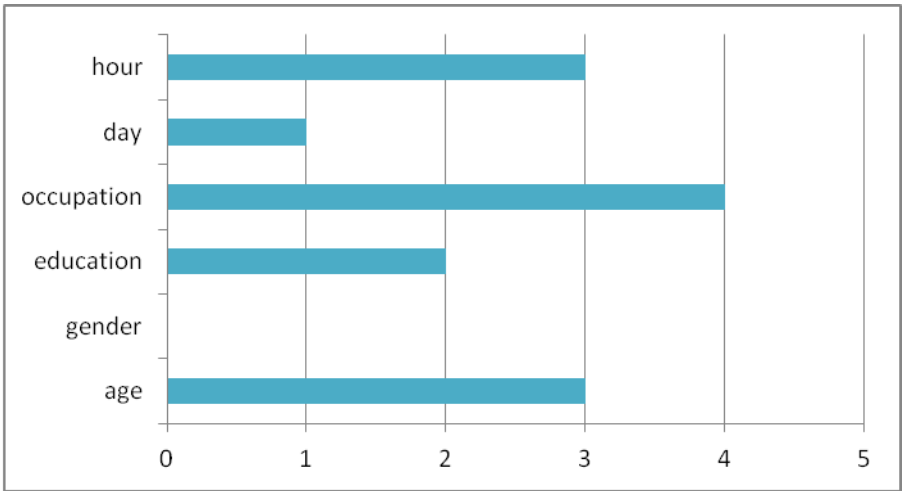


Figure 5: Frequency of attributes appeared in CART

According to these results in Figure 5, the attributes which appeared in the trees are considered as significant. Occupation is the most frequent attribute that appeared in the trees where only the gender was not appeared in any of the tests. Therefore gender is excluded from the NBC inputs. The model is adopted for recommendation prototype system. In the prototype system the NBC is applied to these attributes since attributes which are not appearing in the tree are considered as irrelevant attributes Han & Kamber (2006). To improve the performance of this model and to transform it to learning system, user feedbacks will be included. Following section 3.8 introduces the recommendation approach.

3.8. Recommendation Generation

Classification is useful to construct a user profile especially for the first users, because for such users the system does not know their preferences. This is called new user cold-start problem. In such a case, only their demographic data can be informative to find similar people. Basic similarity measures in recommender systems are cosine similarity, Pearson correlation, Euclidean distance, and NBC (Shih et al., 2011). Chen & He (2009) proposed to calculate similarity between two users by looking at their demographics rather than ratings. Wang & Tan (2011) and Khatri (2012) recently offered a recommendation approach in which they utilized NBC as a similarity measure.

In our study we also perform the similarity calculation by the help of NBC which assigns the user to the other users that have similar demographic and contextual characteristics. In our approach instead of calculating the similarity between two users, the similarity between the target user and other users' is calculated. After the registration, when user logins to the system he / she is asked to specify the day and time preferences. Since the system needs to evolve during time and adapt to user's preferences, we include a feedback of user to the system. For this, we offer first three top genres that the user may like and ask user to evaluate the recommendations related with the user's specified day and time. This best bet recommendation is also known as Top-N recommendation (Cremonesi et al., 2010). The preferences of the users can be affected from various conditions: user moods, accompanying people, day of the week and time affect users' movie genre choice (Nunes & Hu, 2012). Thus when a recommendation of a user is calculated, it is assumed that users tend to behave in a similar way, in similar day and time conditions. Hence, his ranking history is calculated within the same day and time combination.

Since a validation with real users is aimed in this study, due to time limitation, feedback is gathered through ranking but when a real system is launched, instead of rankings customers behaviors can be gathered implicitly by looking with which frequency they watch a particular movie genre. This approach will be useful for the websites which sell tickets of entertainment events and collect only customer's demographic data only in the registration stage. In this type of websites, people do not tend to rate/ rank any item. They only want to buy a ticket. With pursuing customer's buying behavior, website can consider frequent purchased movie genres as feedback of users. Therefore, usage of rankings is symbolic in this study to only increase the weight of a genre which is in the lower level of the results of NBC but actually among the highest preferences of users. The user may like or not like the recommendation, or does not like the ordering that the system offers. If the user does not like either the order of the genres or the genres appearing in the recommendation list, the system asks uses to rank the genres from 1 to 6 and dynamically updates the conditional probabilities of the said user, in other words, the age, education, occupation, gender, day, time preferences of the genre which has the 1st rank. If they like the recommendation, the system does not update anything.

The rankings are saved in the database for the specific users with the specific day and time input. For each new login, the system asks the time and day preferences to the user and if the user gives the same day and time combination, then the system calculates the average of rankings. A user-genre rank matrix is created for each registered user. In this way, next time the user logs in to the system then the system offers users more personalized recommendation by looking both at similar people and his/her own past behavior as well. Personalization can be defined as the adaptation of the offerings of websites to users' preferences and history (Eirinaki & Vazirgiannis, 2003).

The average rankings and the classification results of NBC are combined in the following way:

Consider the posterior probability result for the any genre for the *X* profile as "P (genre | X)" and average ranking of the said genre for user X as "avg (genre rank)", then the combined genre weight will become the following:

$$\text{Weight of Genre}=P(\text{genre}|X)+P(\text{genre}|X)*\frac{6}{\text{avg}(\text{genre rank})}$$

Since we make the user to order genres from 1 to 6 in which value 1 will be the most liked one, with this calculation we increase the value of the classification with the ranks that users give to the corresponding genres. Since the weight is dependent on both the posterior probability of the genre and the rank the specific user assigns to it, this calculation will not make a drastic change in the classification results in order to not dominate the classification result of the NBC with ranks. A simple demonstration is provided in Appendix G. Moreover, we do not prevent serendipitous recommendations. The final weight therefore depends on either the user’s preference behavior or the behavior of the demographic community that the user belongs to, whichever is stronger. Working mechanism of the recommendation is explained also in section 4.1 with an example scenario.

This past behavior is taken into account only dealing with the day and time preferences of previous logins that match with the current login. This way, the system learns the behavior of the user by catching the particular user behavior pattern in particular day and time combinations. Otherwise, the system provides the recommendation based on just the NBC. To sum up, the user takes recommendation based on only NBC, he does not provide no more than two same day and time preferences, else the system takes the average of same time and day preference pattern and provides the recommendation not only based on NBC but also user’s day and time and past movie ranking decisions.

CHAPTER 4

PROTOTYPE RECOMMENDER SYSTEM DEVELOPMENT AND VALIDATION

In this chapter, the overview of the implemented prototype along with its architecture and graphical user interfaces is presented. The experimentation details with real users are also provided.

4.1. Aim and scope

According to the results, it is concluded that designating the CART tree as a feature selection method for NBC increased the number of correctly classified instances. Even though, the accuracy results are not very encouraging, when the improvements are compared with previous studies conducted in different domains by (Ratanamahatana, & Gunopulos, 2003; Ting et al., 2011), improvements can be accepted as satisfactory. The discussion about this issue is provided in Chapter 5. To make a validation, a prototype system is designed and developed. Since a personalized recommender system is aimed to be built, there is a need to collect user feedback consequently changing individual preferences. It is natural that, choice of people changes over time and not all people having same demographic profile have the same tastes (Wang & Zhou, 2012). Therefore, in addition to users' demographic profile and day and time choices, users' feedbacks and genre preferences were also included to the scope of prototype. In addition to the prototype a small questionnaire is performed with users of the prototype system. The following sections introduce the details of the validation.

4.2. Prototype Recommender System Overview

This section presents the system architecture, database and system's working mechanism. A typical scenario for user is also provided.

4.2.1. System Architecture and Development Details

For the web based prototype 3- tier architecture is adopted. These types of architectures have 3 layers: presentation layer, business layer and data layer. In the presentation layer, browser and GUI application exit. This layer takes commands from the user and transmits them to the business layer. Similarly, it takes data from the business layer and presents them to the user. For the case of this study, the user provides data such as their demographics and day / time options as well as feedbacks such as movie genre rankings to the presentation layer and sees the recommendation results from it.

The business layer contains the web server. The layer is responsible for doing the business such as processing the demographic data coming from the presentation layer, and transmitting it to the data layer.

The third layer is the data layer, which has the ability of accessing the database and performing some actions on it. These actions include reading, writing, and updating the database. For the movie recommendation problem, a new user is inserted to the database through this layer or a conditional probability of the said profile is updated again through this layer. 3-tier architecture of the prototype is illustrated in Figure 6.

Web based prototype is developed using Java Servlet for Server Programming and Java Server Page (JSP) for web interfaces. Servlet technology provides to create dynamic web pages. Servlet applications are java classes which work on server and provide results according to the data gathered from the users. JSP are used to create web pages based on HTML in which Java Servlets can work.

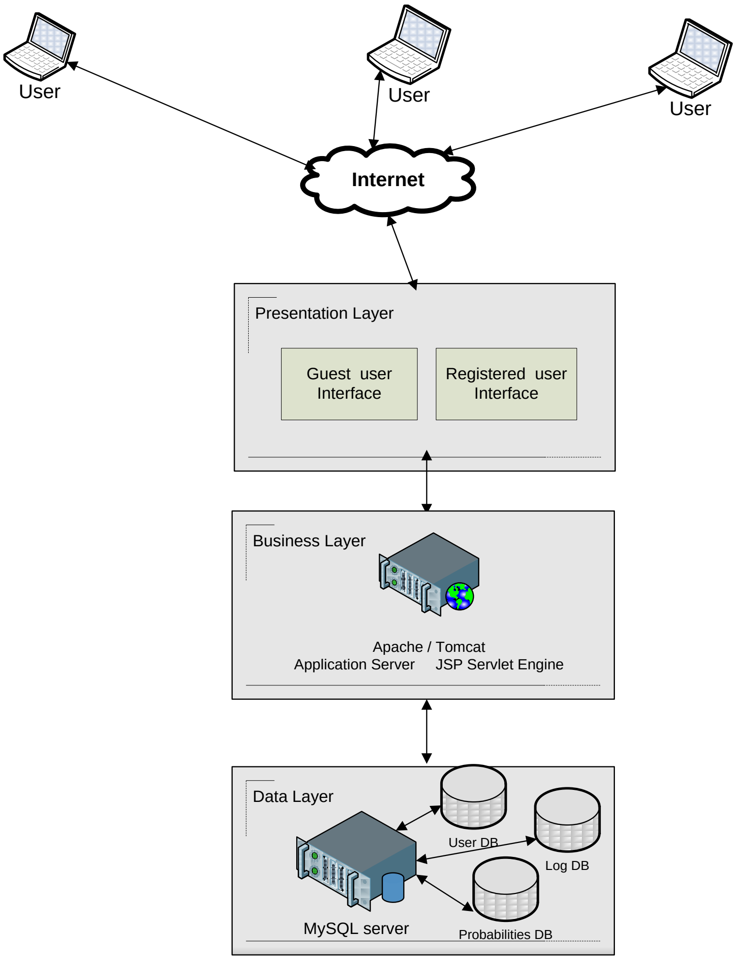


Figure 6: Three-tier System Architecture

4.2.2. Database

In the system, there are three different data sources. The first one is the probability data, in which movie genre classes and user attributes' conditional probabilities exist along with the classes' prior probabilities of original purchase record dataset. At previous stages, random sub sampling was used and the dataset were divided into the training and testing sets. At this stage it is not possible to divide the data into two subsets such as training and testing set since according to Stritt et al. (2007) it is only possible to consider the data from the past to train the model. "In real life, only data from the past can be taken into account for training because no data from the future is available". This argument explains the reason why original purchase data is used for the probability data in the prototype system.

The Naïve Bayesian classification will be performed with the help of the probabilities existing in that database. The second data is the user data, which keeps the demographic data of the registered users along with their usernames and passwords. Third data source is the user log data which keeps the different login information of the users. Each user may use the system several times. Moreover, their time and day preferences as well as movie genre prioritizations may change over the time. These changing behaviors are kept in this database. The second and third data will be collected with the experiment. Each data source corresponds to a table in the database. As a database management system, MYSQL Database Management System is used.

4.2.3. System's Working Mechanism

In the prototype system, a learning recommender system is aimed to be built. This learning system adapts according to the customer feedback. In addition to the customer feedback in the developed system, users are asked to rank, in other words, to re-order the movie genres that were offered by the system. The flow of the study is given as follows:

There are two types of users in this prototype system: guest users and registered users. The use case diagram given in Figure 7 below shows the main functionalities of the system from user perspective.

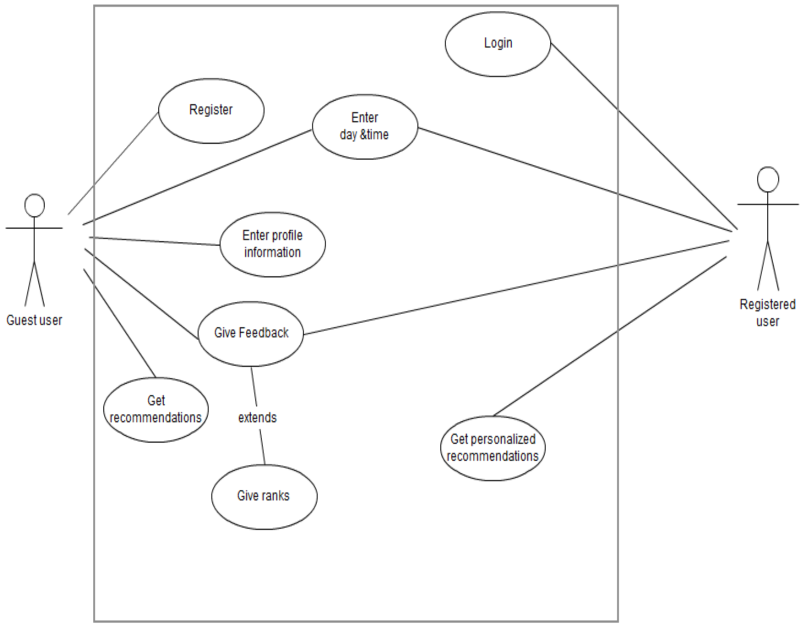


Figure 7: Use Case Diagram

The following activity diagram given in Figure 8 shows the registration procedure of a new user to the system. User provides their demographic details to the system for registration. The activity diagram given in Figure 9 shows the receiving recommendation and giving feedback procedure of the user.

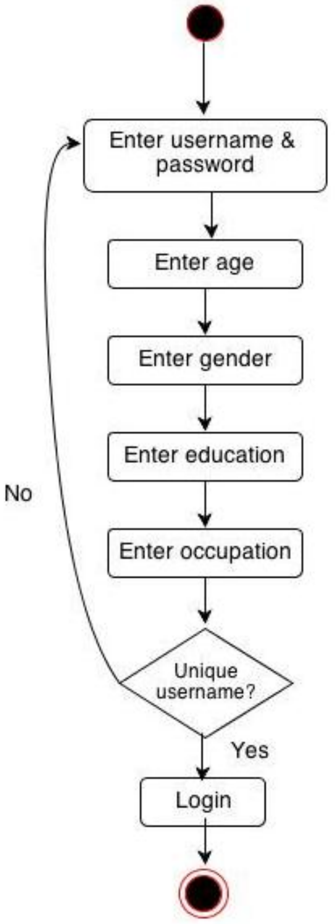
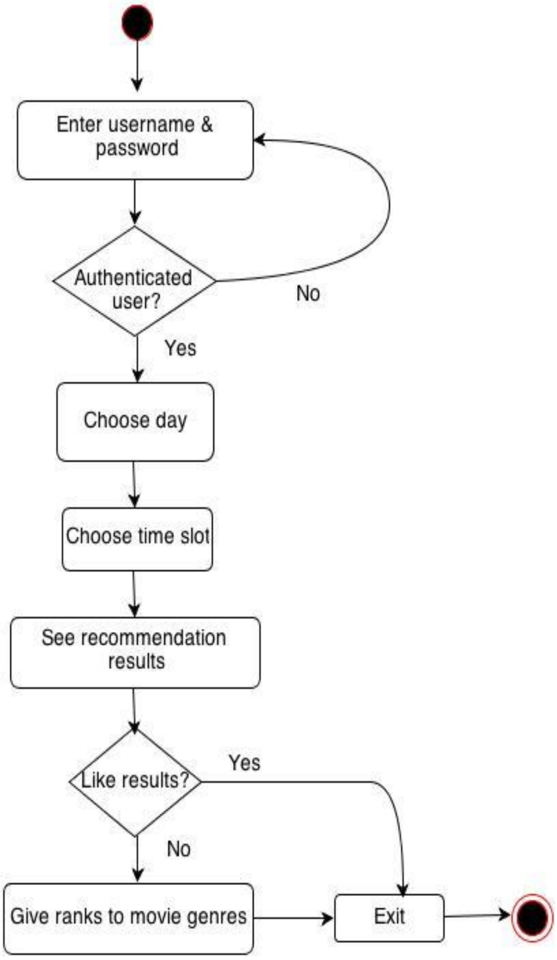


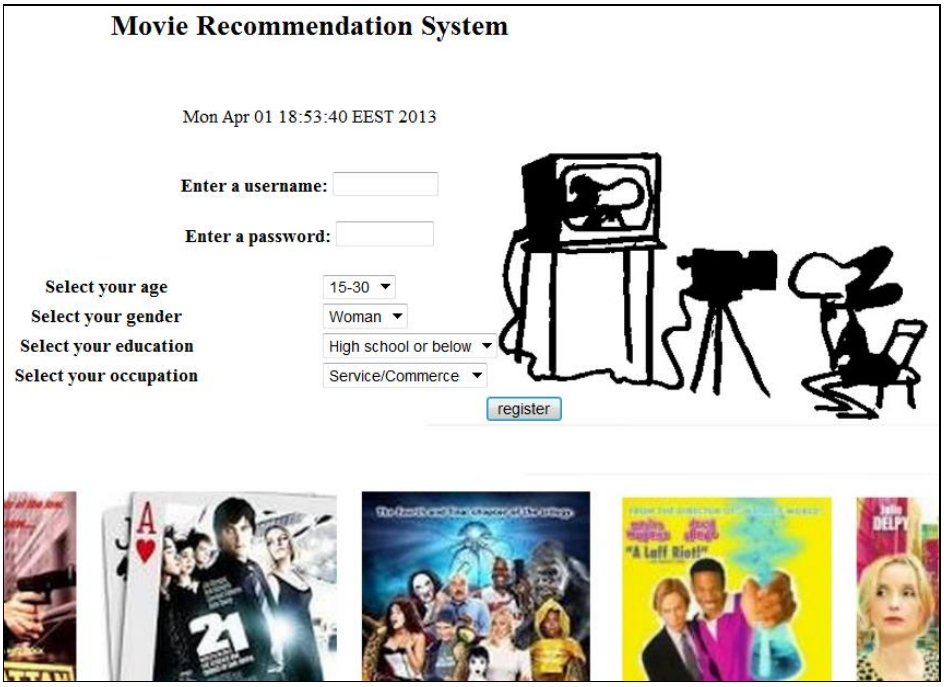
Figure 8: New user registration



Enter username & password
Authenticated user?
No
Yes
Choose day
Choose time slot
See recommendation results
Like results?
Yes
No
Give ranks to movie genres
Exit

Figure 9: Give feedback

"Guest users" can use the system anonymously and their log information is not kept by the system. They can also provide their demographic details to the system and the system provides them three movie genre recommendations with respect to the profile that they belong to. If they agree with the recommendations presented by the system, the system does not update anything. However, if they do not agree with the recommendation results, then their feedback is taken. The probability table is updated according to the movie genre that they prefer the most. Since the probability table has the profile and genre frequencies, only the first choice of the guest user is taken into account. The second type of users of the system is “Registered users”. While registering they provide their demographic data to the system, and these data are kept in the user database. Figure 10 shows the user interface of the registration of a new user.



Movie Recommendation System

Mon Apr 01 18:53:40 EEST 2013

Enter a username:

Enter a password:

Select your age 15-30

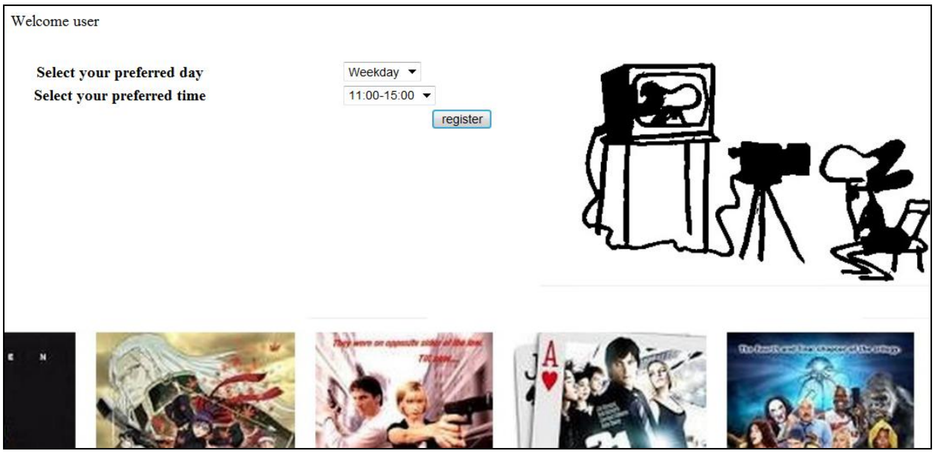
Select your gender Woman

Select your education High school or below

Select your occupation Service/Commerce

Figure 10: Registration Screen

After registering, users can login to the system several times. The system recognizes these users. Their log details are kept for each login. The system gives the option to the user to choose the day and time preferences in each login. Figure 11 illustrates the day and time preference entrance of a user.



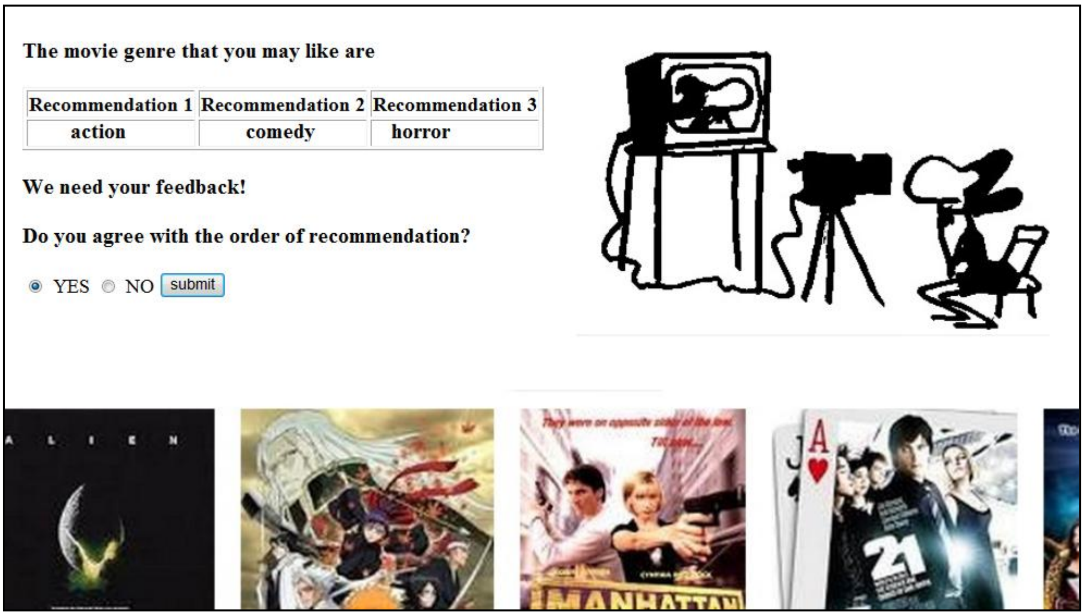
Welcome user

Select your preferred day Weekday

Select your preferred time 11:00-15:00

Figure 11: Choosing day and time preferences user interface

This is done to evolve the system with the time and make an intelligent system which learns the user behavior over the time. The system presents three movie genre recommendations to the users which are the top 3 movie genres that can be appropriate to him or her. When the recommendations are presented to user, the user has two alternatives: agreeing with the genres and order of them, that are recommended to him/her or not agreeing. Figure 12 shows the feedback of users



The movie genre that you may like are

Recommendation 1	Recommendation 2	Recommendation 3
action	comedy	horror

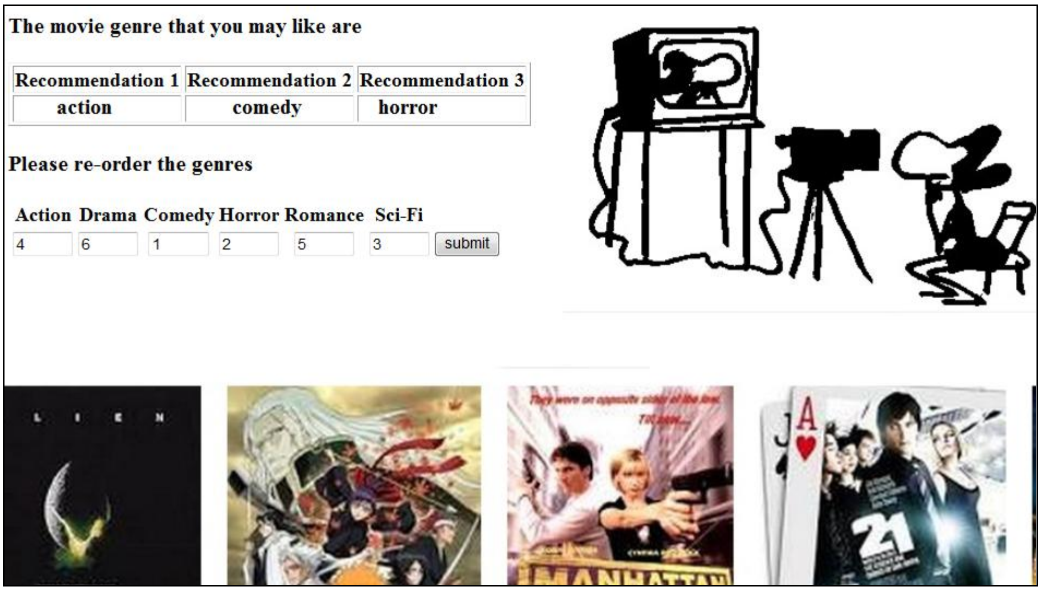
We need your feedback!

Do you agree with the order of recommendation?

● YES ○ NO submit

Figure 12: Recommendation results and user feedbacks screen

If the users agree with the recommendation, the system assigns rank 1 to the first, rank 2 to the second, rank 3 to the third recommendation; and assigns rank 4 for the remaining movie genres which were not recommended by the system to the user. In this stage there is no update in the probability database. If the user does not agree with the recommendation results, then the system provides the user a list of movie genres and text box in which the user can input the ranks of the movie genres according to his preferences. These rankings and user's each day and time preferences are kept in the log database. The probability table is updated according to the 1st genre that the user prefers. Figure 13 shows the user interface for this activity.



The movie genre that you may like are

Recommendation 1	Recommendation 2	Recommendation 3
action	comedy	horror

Please re-order the genres

Action	Drama	Comedy	Horror	Romance	Sci-Fi
4	6	1	2	5	3

Figure 13: Provide ranking screen

The recommendation working mechanism can be explained in the following way. If it is the first login of the user, then the system gives a recommendation to that user, benefiting from the other users who have a similar profile with the help of DT enhanced NBC classification. However, if the user logs in to the system for the second time, then the recommendations of the system are not based on only the NBC but also his / her past preferences along with the average ranking information.

4.2.4. A Typical Scenario

Consider the scenario for a user "X" with the following attributes:

Age	Gender	Education level	Occupation
26 years old	Male	University	health sector

Suppose the log details of the "X" are:

Table 5: User- genre rank matrix

Date	user	Day	hour	Comedy	drama	romance	sci-fi	action	horror
08.03.2013 13:42	X	weekend	18:00	3	2	5	1	4	6
16.03.2013 16:30	X	weekend	18:00	1	4	5	2	3	6

Since in the tests conducted in section 3.7.3 it was found that gender has no effect in the classification according to CART, the movie classes weights are calculated without considering the gender of users.

P (comedy | X) = P (university | comedy) x P (below30 | comedy)
x P (health | comedy) x P (weekend | comedy)
x P (evening | comedy)
P (comedy | X) = P (comedy | X) + (P (comedy | X) x (6/average (comedy)))
P (comedy | X) = P (comedy | X) + (P (comedy | X) x (6/ (3+1)))

Other classes' weights are calculated exactly in the same manner. In this way both the NBC and simple moving average past movie genre choice have some effects, in the recommendation process. Not a single method is dominant in every case.

4.3. Experimentation and Validation

To validate the system, real user data is collected. While choosing subject people, the frequencies gathered from the CART decision tree analysis performed in section 3.7.3 is taken into consideration. According to this analysis, most frequent demographic attribute in the original purchase record dataset was found as occupation. For this reason, subjects were selected to cover all occupation categories, having many different age and education combinations. 28 people experimented with the system. In addition to the system experience, a small questionnaire was applied to the users, containing questions about movie choices. The questionnaire can be found in Appendix D. From 28 users, 10 of them logged to the system just once and 18 of them logged two or more times. This way, 52 user log records were obtained. The following section introduces the results of validation. The demographic distribution of the participant users are provided in Appendix B.

CHAPTER 5

RESULTS AND DISCUSSION

Results and discussion of the experiments conducted at this study are introduced in this chapter. First some explanations related to the comparison of traditional NBC and DT enhanced NBC is given. Later, some results and discussions are provided about the evaluation of the prototype.

5.1. Results and Discussion for Comparison of different classifiers used in this study

In our experiments that are conducted both with CHAID and CART, although the improvements are not in a major scale, it is proven that DT enhanced NBC, in other words usage of DT for the feature selection of NBC reached to better classification results than standard NBC. It is seen that, in the previous studies conducted in different domains where DT is used as a feature selection method for the NBC by Ratanamahatana & Gunopulos (2003) the improvements were also in a minor scale; in various datasets they used improvement which was ranging among numbers at least 1.1% to at most 7.9 %. A similar study was conducted by Ting et al. (2011) in which they used chi-square statistics for the feature selection of NBC in the documents classification domain and accuracy improvement gathered was 0.1 %. Ting et al. (2011) pointed out that even though the improvement is not very encouraging; the feature selection approach yielded a better classification when it is compared to traditional NBC technique. Therefore, we can say that DT is also successful in movie classification domain as a feature selection method for NBC.

Moreover, it should not be forgotten that the system is self-learning and therefore the classification performances will be updated and improved over time. However, we also make some comments about the dataset that we think has effects on the classifiers performance. Konstan & Riedl (2012) pointed out this issue that data can be considered as one of the most significant challenge in recommender systems research by being wrong, corrupted and noisy. Moreover, Du (2010) reported that data quality affects directly the data mining results and emphasized that it is difficult to measure the quality of data from erroneous entry perspective. Data set used in this study is a real dataset which was gathered from an online electronic ticket company of entertainment. This online company provides tickets from various entertainment activities ranging from theatres, cinemas, sports, events, concerts, shows and travels. The reasons of the low accuracy of the tests conducted using this data can be the following:

- Since it is a historical data, it may not be really random and may be inclined to the people who are using the internet, in other words, it may represent specific group of people and not reflect the whole population.
- It is collected for a specific time interval in which there is no proper distribution of different movie genres. We found that the action type movies were nine times more records than other movie genres; to not cause bias in the sampling stage we took the same number of records for each genre.
- It is not collected by aiming just the cinema/ movie domain therefore only the genre attributes for the movies were available.
- It reflects people who live in cities where the cinema halls allow ticket buying through the internet.
- Since it is an online record, we do not trust the correctness. The values in the data may be filled incorrectly. It is stated by Du (2010) that data collected in a real environment can be different from the one collected in a controlled environment.
- Even though some pre-processing activities were conducted, it could be possible that some hidden problems in the data were overlooked. For example, many inconsistencies are observed in the dataset especially with age

and education level, these were eliminated but the inconsistencies which are not so obvious may still be in the dataset.

All these, can be considered as the reasons for the low accuracy of the initial test results.

With the help of CART, it is aimed to eliminate less significant attributes from the Naïve Bayesian Classifier Calculation. According to CART decision tree, only the gender attribute is eliminated from the attribute set even though gender is generally considered as a factor in recommender system studies (Ujjin & Bentley, 2002; Sobecki 2006; Wei et al., 2007; Chen & He, 2009; Chikhaoui et al., 2011) in the literature. The reason behind this may be for this specific data set, gender may not have an important role because since dataset is made up of the records of an online ticket selling site and the person who purchased the ticket it not always the person who will watch the movie. To put in another way, wife/ husband may book the tickets for each other or children may book the tickets for their parents and vice-versa. The reason can consequently be the wrong filling of the gender information in the system. Moreover, we already know that people are not willing to give their private data too. Montaner et al. (2003) summarize this issue by claiming that due to privacy concerns, people do not want to engage with the websites, consequently no data or wrong data are provided generally.

In the prototype system, users were assigned to classes, firstly according to their demographic profile as well as day and time preferences using NBC as similarity measure. Later, the feedbacks of the users are gathered; whether they were satisfied with the recommendation provided to them. In addition to this feedback, users' movie ranking preferences were taken into account. The reason behind this approach is firstly to distinguish users from the community that they were assigned at the first stage and to provide more personalized recommendations. In each login, the user faced the recommendation according to demographic community that he/she belongs to as well as the average of his/her own past preferences. Ricci et al. (2011) point out the issue that it is not a good idea to consider only the demographic data of people while making the recommendations. However, they also summarize the importance of the demographics in this process by stating some strength of using demographic data such as it prevents cold-start problem- even if no data exists about the past behavior of the user, the system can classify users to movie genres by considering other people demographic profiles; it is independent of the domain that is used and it may give serendipitous recommendations. As an example for the serendipity, we observed one case in which a user from the health sector group- a doctor is not satisfied with the recommendation given to her. According to the demographic profile that user belongs to the probability of assigning to a horror movie has not appeared among the recommendation results presented by the system. Here, the user was faced with other movie genres that other people similar to her profile chooses most, as a serendipitous result. Moreover, in future, usage with the help of ranking, the movie genre which is in the lower level according to the profile this user belongs to, would increase to the top-3 levels.

5.2. Results and Discussion for Prototype and Experiment

Although in this study, people are supposed to rank the movie genres according to their preferences, it should be said that this is not a ranking problem actually. Because people may not be able to express precisely which movie genre they like more or less than the other genre. They can like both genre the same. As stated in the previous section 5.1, this ranking assumption is made only to get feedbacks from user to improve the recommendation part.

This feedback offers a chance to a movie genre which appears in the lower levels in the recommendation results according to NBC but at the same time, which is one of the most preferred genres of a specific user, to move the upper levels with the help of recommendations. This way, system evolves from generalizing users' preferences state to personalizing state. Therefore, while evaluating the validation results, the evaluation is grouped into two categories:

1. Evaluating the accuracy of generated recommendations without considering the order of the recommendation
2. Evaluating the accuracy of generated recommendations considering the order of the recommendation

The performance of the classification is generally measured by evaluating its accuracy (Dunham, 2003). Classification accuracy metrics examine the decision making ability of the recommendation algorithms (Schröder et al., 2011). Therefore, for the performance evaluation, the accuracy measure is adopted as stated in section 3.6. Confusion matrix is a standard tool that is used to show the performance of a classifier (Sharma & Mukherjee, 2012). It shows the accuracy of classifiers by creating an n x n matrix where n is the number of classes in the classification problem. The values situated in the diagonal line of the matrix shows the number of records classified correctly and remaining parts of the matrices shows the number of misclassifications. A generalized version of the confusion matrix is given for a binary classification problem in the following Table 6.

Table 6: General Confusion Matrix for binary classification

Actual Classes	Classified as A	Classified as B
A	True positive (tp)	False negative (fn)
B	False positive (fp)	True Negative (tn)

The evaluation of classification and recommender system performances is usually performed by using some metrics such as accuracy, precision and recall. MAE (Mean Absolute Error) is one of the most popular one. However, McLaughlin & Herlocker (2004) pointed out that recall and precision are better classifiers than MAE. Cremonesi et al. (2010) proposed to measure the recommendations with accuracy metrics such as recall and precision. The general definitions of these metrics are given by Sokolova & Lapalme (2009) as; accuracy, the overall success of the classifier is defined as total number of correct classified examples over total number of classifications. Precision is defined as the ratio of the number of correctly classified positive examples over the total number of examples that are classified as positives. Recall is defined as the number of correctly classified positive examples over the total number of positive examples available in the dataset. Even though in the literature, much of the studies these metrics were calculated for binary classification problems, it is found in a similar study conducted by Ting et al. (2011) where the performance of NBC for document classification was investigated; recall and precision metrics were adopted for the performance evaluation of classification of four classes. These two metrics can directly be calculated from the confusion matrix. The formula of these metrics for binary classification problems are given below in the following Table 7:

Table 7: Evaluation Metrics formulas

Accuracy	Precision	Recall
$Accuracy=\frac{(tp+tn)}{(tp+tn+fp+fn)}$	$Precision=\frac{tp}{tp+fp}$	$Recall=\frac{tp}{tp+fn}$

As stated before, the available studies show example evaluation usually for the binary classification which is not the case in this study. In this study, it is intended to classify people to a movie genre among six different movie genres. This type of classification is called multi-class classification. Sokolova & Lapalme (2009) in their studies where a systematic analysis of performance measures for classification task is presented, provide a formal measurement for the multi-class classification problem. The precision, recall and the accuracy are given in the following table for the multi-class classification problems:

Table 8: Metrics for multi-class classification problems

Average Accuracy	Precision	Recall
$Average\ Accuracy=\frac{\sum_{i=1}^{l}\frac{tp_i+tn_i}{tp_i+tn_i+fp_i+fn_i}}{l}$	$Precision=\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fp_i}}{l}$	$Recall=\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fn_i}}{l}$

While evaluating the performance of the prototype recommender system, for the accuracy measure the second guess heuristic approach given by Han & Kamber (2006) is adopted. In this heuristic, it is said that if a tuple is likely to belong not only to one class but also to more than one class than the accuracy of it can be measured by considering the prediction result as correct if it matches with the first and second most probable classes. We extend this heuristic by considering the third most probable class to be consistent with our case. However, Han & Kamber (2006) also mention that there is still no complete solution for multi class classification problems. Amatriain et al. (2011) also supports this unavailability of the evaluation. Beside, Sampat et al. (2009) in their study worked for the evaluation of three class classification problem and also mentioned that there is still no widely accepted methodology for assessing the performance of the multiclass problems.

Hence, it is challenging to evaluate the performance of a classifier which has more than two classes, in other words, multiclass classification problems. Therefore, while constructing the confusion matrix; we considered a movie genre which is in the top three actual preferences of a user, to be correctly classified if it appears in one of the three recommendation results of the system.

Result 1: Evaluating the accuracy of generated recommendations without considering the order of the recommendations

The confusion matrix of the classification without considering the order of a genre is given in the Table 9. In this table, if a genre appears both in the actual preferences and recommended genres of the users, it is counted as correctly classified, in other words, successful; independent of its rank in the recommendation results. This type of measurement in which counting correct the matching of genre independent of its rank also suggested by Sokolova & Lapalme (2009)

Table 9: Confusion Matrix for classification of top 3 level genres

Actual Genres	Predicted Genres					
	Comedy	Drama	Horror	Romance	Action	Sci-fi
Comedy	25	2	6		1	3
Drama		28	2		1	
Horror		4	11	1		
Romance	1			21		3
Action	1	1	1	3	14	1
Sci-fi	2	1	1	3	3	16

According to the average accuracy formula given in Table 8, it is found that the average accuracy for overall classification is 74 %. The precision and recall results of top 3 level genres from the confusion matrix in Table 9, without considering the order are given as presented in Table 10.

Table 10: Precision and Recall results for overall classification of top 3 levels without considering the order

Genre	Precision	Recall
Comedy	0.86	0.68
Drama	0.78	0.90
Horror	0.52	0.69
Romance	0.75	0.84
Action	0.74	0.67
Science-fiction	0.70	0.62
Average	0.72	0.73

According to the average precision and recall results presented in Table 10, the average recall results is higher than a similar study which offered Top 20 movie recommendation by Cremonesi et al. (2010). To make a comparison recall results of only top three recommendations are taken into account. Their recall results change from approximately 0.10 to 0.65. Similarly, a recommender systems study conducted by Bobadilla et al. (2011) exploiting from neural networks to avoid cold-start problem, presented from top 2 to top 20 recommendations. To compare the precision and recall results obtained in our study with that study, again the results for top three recommendations are observed and it is seen that the precision results were changing between 0.45 and 0.60 and recall results between 0.20 and 0.50. Since top 3 levels movie genres are offered to users over 6 movie genres, it can be normal that the recall results are higher than other studies. Even though the technique adopted is different, it can be said that the results obtained in that study are within acceptable levels. As a summary, high recall means most of the relevant results are returned by the algorithm where high precision means the algorithm returned considerably more relevant results than the irrelevant ones (Wikipedia, 2013).

Result 2: Evaluating the accuracy of generated recommendations considering the order of the recommendations

When the order of the movie genres is taken into account, the following Table 11 is obtained by showing the accuracy results of all the 52 records.

Table 11: Accuracy results of recommendations with respect to their positions

Total records	Correct ordering	First genre match	Second genre match	Third genre match
100 %	36.53 %	57.7 %	51.9 %	49.0 %

According to Table 11, it is concluded that

- 36.53 % of the records got the recommendations in the correct order.
- For 57.7 % of the records, the actual first preferred genres match with the genre recommended as first by the system.
- For 51.9 % of the records, actual and recommended results for second genre match.
- Similarly, for 49.0 % of the records, actual and recommended results for third genres match.

In a study conducted by Casali et al. (2008) where an agent based recommender system is studied in the tourism domain, researchers performed a validation where they used the feedbacks of users in terms of ranking. Users' preferences and restriction were collected explicitly through online forms. Users were supposed to provide their preferences and restrictions by giving numbers from 1 to 10. Researchers used approximately the same number of users (30) and logs (52) in their validation set and they found that the correct ordering accuracy result of the above mentioned study was 40.4 % which is very similar to the results obtained in this study. This finding encourages our work, since the mentioned study also considered the user preferences and collected these preferences explicitly as ranks but for another domain which is tourism domain exploiting another approach that is the multi-agents models. Moreover, recall and precisions for three ranks are also calculated and presented in Appendix A.

<u>Result 3:</u> Measuring the learning ability of the system

<u>Result 3.1:</u> Considering the correctness of the ordering of recommendations:

In the experiment conducted, 28 users were offered recommendations according to the NBC in their first log. To investigate the learning ability of the system between the first log and the remaining logs, the improvement of giving correct recommendations are assessed in terms of accuracy. The assessment is divided into two stages such as first log and other logs. First log is important because it offers recommendations with respect to NBC and DT combined technique and next logs depend on both NBC and DT and also users' feedbacks. Hence, learning can be monitored by looking at the records to which correct recommendations were provided. From 28 users, 10 of them logged only once and remaining 18 of them logged two or more times. Here are some results:

- Since there are 28 people using the system and each user is logged at least once, 28 recommendations are generated using Naive Bayesian Classifier. 6 of 28 recommendations presented correct movie genres in the correct order. Therefore the accuracy is 21.40 %.
- Later, 24 recommendations were done either in the second log or next logs. These recommendations were provided with respect to both NBC and user feedbacks. 13 of 24 recommendations presented correct movie genres in the correct order. The accuracy is found as 54.16 %.
- The improvement of the system is 32.76 %.
- Therefore, the overall accuracy of the learning recommender system over time is reached to 19 users having correct movie genres with correct ordering. In other words, in the end of the experiment 67,8 % of the users received correct recommendations over time.

Result 3.2: Not considering the correctness of the ordering recommendations:

- At first log, 9 of 28 recommendations, 32.14% were found to be correct, without looking at the position of the recommended movie genre appeared.
- At second and third log, 14 of 24 recommendations 58.33 % found to be correct, without looking at the position of the recommended movie genre appeared.
- Here, the improvement is found as 26.19.
- At the end of experiment 20 of 28 people, 71.42%, received correct movie genres not considering the position that the genres appeared.

With these results, it can be seen that a user whose preference history is not available still can receive recommendation with the help of similarity calculation through NBC. As stated by Aimeur et al (2006), Ricci et al. (2011), demographic data helps resolve cold-start problem that is available in CF and CBF. Additionally, it can be said that NBC is also a useful and easy similarity calculation technique for demographic similarity calculation like other methods Term Frequency (TF) and Nearest Neighbor Set previously used by Chen & Le (2009). However, it is stated that the accuracy of their algorithm is dependent of the number of *‘N’* nearest other users. Different from them, there is no need for selecting *‘N’* number of nearest neighbor users in our study, since our system tries to correlate between one active user and other users. Besides, while providing a recommendation, in addition to looking at user similarities, users’ past behavior is taken into account. However, as stated by Vozalis & Margaritis (2004) it should be noted that demographic data is not sufficient alone to make recommendations and it gives more improved results when combined with other techniques that is learning from users’ past behavior in this study. Still, with learning, it can achieve good results.

In addition to the prototype experiment, a small questionnaire is applied to the users. The questionnaire can be found in Appendix D. In this experiment it is asked to user to rank the features such as actor, director, genre, language, award, year from the most affecting to select a movie to the least affecting. People gave rank 1 to the features genre, actor and director. According to the results, it is found that genre is the most important attribute considered for people to select a movie since 19 people over 28 gave rank 1 to the genre attribute, actors and directors follows genre. Number of people who gave rank 1 to the genre, actor and directors are given in Figure 14.

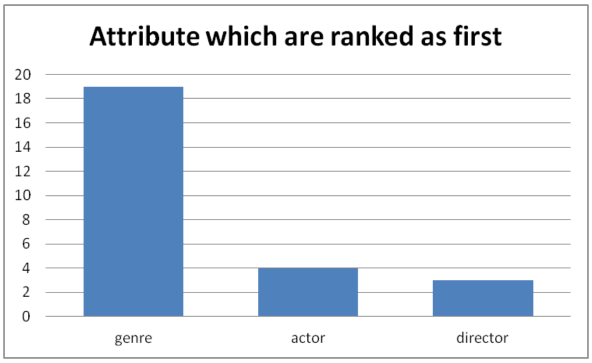


Figure 14: Distribution of movie attributes that received 1st rank

This fact supported our assumption at the beginning of the study that people choose movies by looking at the genres. The least important features are found as year, award and language of the movies. The overall ranking of the features can be seen in Figure 15.

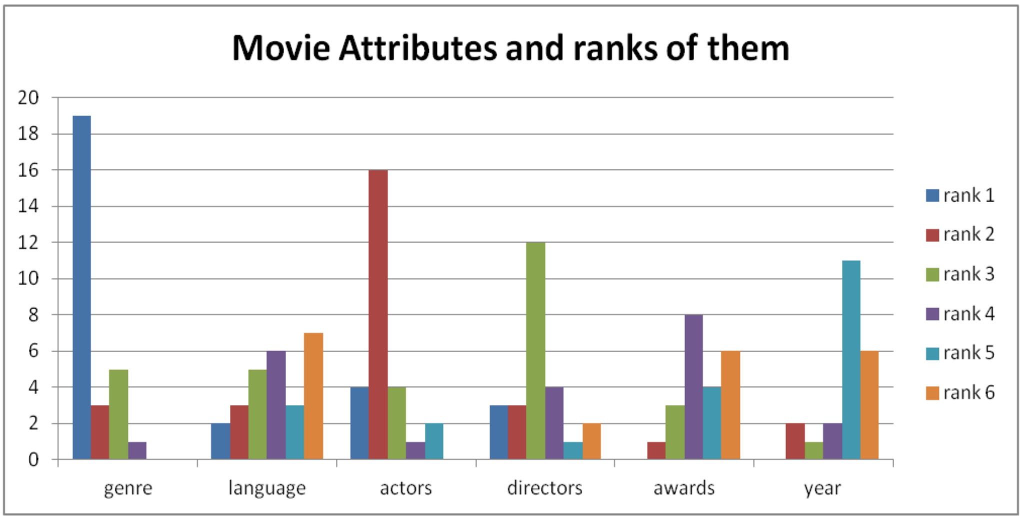


Figure 15: Distribution of rankings among movie attributes

According to the results 19 over 28 people change the movie genre with respect to the accompanying people. In addition to this, the movie genre changes also the identity of the accompanying people such as family or friends. The following Figure 16 shows how the distribution of movie genre preference changes over accompanying people.

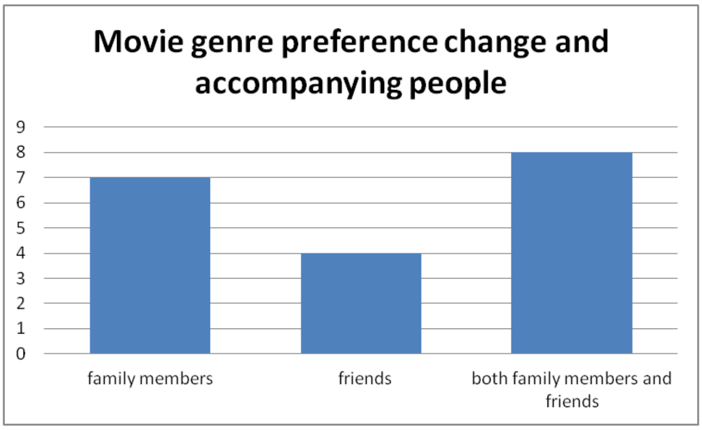


Figure 16: Distribution of movie genres changing according to the accompanying people

From Figure 16 it can be observed that family members such as (parents, children) affects the movie genre choices. For this reason, the historical purchase data for a movie cannot be informative about the user own preference. User preferences should be taken into account and must be collected either explicitly or implicitly.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1. Conclusion

Recommender systems are studied by many researchers from various aspects. It is an important research area which has both economical and productive sides both for users and content providers. Users can easily reach the information that is related to them among a large pool of information; they can save time. Companies may profit from providing good recommendations to their customers. With this way they can improve both the user satisfaction and customer loyalty. A hybrid approach exploiting the strengths of two techniques NBC and DT for the classification to build a personalized demographic recommender system constitutes the focus of this study. The conditional independence assumption of NBC that can be considered as an advantage of NBC which can be useful to handle missing attributes in a classification process, can easily become a disadvantage of the technique in the cases where there exist dependencies among attributes. To overcome this issue, decision trees are suggested by the researchers to select the significant attributes while eliminating less significant and / or redundant attributes. The success of decision trees as a feature selection method for Naïve Bayesian Classifier was demonstrated in various domains except movie domain.

Moreover, to best of our knowledge, no recommender system is built with this technique. In this study, firstly the success of DT enhanced NBC is investigated while classifying users according to their demographic profile to the movie genres. It is found that DT as feature selection technique for NBC outperforms the standard NBC in this domain too. The improvements were not big-scale but the reason behind this can be the dataset used for which we do not know the correctness.

Secondly, after demonstrating the performance of DT enhanced NBC in this domain, a recommender system is aimed to be built. For this aim, a personalized recommendation strategy is proposed considering both NBC approach and movie genres preferences of users as ranking of movies. When including user feedbacks, the aim was not to change the NBC results drastically but to give a personalization to the system. For this reason, a recommendation formula is offered to combine NBC and user past ranks. With this way, the system both considered demographically similar other users and user's self- preferences while making recommendation.

A validation experiment was conducted with 28 people and it is found that although, the validations is done exploiting the previous dataset that we are unsure of its correctness, for the prior probability calculation, it gave encouraging results. The system won a learning ability not dominated by rankings but in an evolutionary manner. Since the system is learning over time, we expect that the accuracy will increase over time. We can conclude that DT can be used as a feature selection method for NBC as classification component of a recommender system. According to the results of a short questionnaire it is seen that people mostly chose a movie by looking primarily to its genre. Their genre choices are affected by the accompanying people. This study mainly referred to the following items:

- NBC was used to classify people according to their demographics and day and time choices.
- Result findings were less accurate.
- Whether combining NBC with other technique can improve accuracy was investigated.
- A model was proposed; DT that was used for feature selection of NBC improved the accuracy and results were found to be better.
- The model was refined by considering the behavior of each user. The average user behavior in the same circumstances such as same day and time combination were taken into consideration.
- The approach adopted in this study was different from the existing approaches, by utilizing DT for feature selection technique for NBC to find demographic similarity between the users and by combining it with the user's behavior.

- In the literature the similarity is taken into account but in this study similarity is found with NBC and DT. With this way a recommendation is provided to users even though there is no feedback or past behavior in the system.
- A personalization is achieved by taking users' behavior into consideration.

6.2. Limitations

This study is conducted by using the historical records of an electronic ticket company of entertainment as dataset. This online company provides tickets from various entertainment activities ranging from theatres, cinemas, sports, events, concerts, shows and travels. The data collected is a log data of customer purchases ranging from music, theatre, movie to the travel domain. The scope of the study was restricted to the movie recommendation domain. Therefore it is a sparse data in terms of attributes for the movies. Especially, even the most general attributes related to the movies were not available in the dataset. Only the genre attribute was informative for the movie. Also, the data set reflects the customer behavior for a limited time interval. Therefore the evolution of the customer over the time cannot be observed.

Moreover, since the aim of the company is to collect this data, just to record the user portfolio of the company, no data like customer feedback, rating etc is gathered.

Even for the demographic information of users only age, gender, education level, occupation and place information were available. Hence, if more data could be gathered over the time, or customers' feedback could be collected, the results of the study would be more informative.

Besides, for the validation part, 28 people is tested with the final system, due to time limitation of the evolution system over a long time cannot be observed.

In the system being proposed, a learning recommender system is aimed to be built. When a recommendation is provided to the user, system asks the user whether he/she is satisfied from the given recommendation and updates the conditional probabilities according to the users' feedback, therefore a learning mechanism works for this system. For this aim, as it is mentioned in the introduction, a hybrid of NBC and DT approach is adopted. However only the NBC part of the system is built as learnable but the DT part is static. The reason behind this is that, the DT is just used to select the features that are significant and that will be input to the NBC.

For the classification, additionally, with the analysis done by changing sample size it is concluded that the branching structure of DT and nodes appearing in the tree changes very hardly thus it requires many examples. Not a significant change is observed in the tree with the small changes of sample size, it is almost the same with the small increases in the samples. Since this is not a longitudinal study, it is not possible to test the system with a high number of real data. Therefore, the learning ability is added only to the NBC part of the system.

Another limitation is inherent from the dataset which is collected through the web. It is not possible to know or discover whether the demographic data provided by the user is really correct. People may fill the blanks with unrealistic data such as a high school student may choose his/ her education level as university for example.

Lastly, for the movie genres six classes were chosen. These are action, comedy, drama, horror, romance and science-fiction. The reason behind this choice is that these types are considered the main-stream, major movie genres and in the dataset we used they were the available movie genres.

6.3. Future Work

In this study, the learning ability is given to the NBC part of the system and DT part is remained static due to the reasons presented in the limitation part of this report. However as a future work if the study can become a longitudinal research and real data can be collected with a long interval of time, decision trees can be built periodically and the features of NBC can be updated accordingly. This way, the long-term performance of the recommender system can be monitored. Since NBC is a flexible technique for missing attributes, tests with missing attributes can be done and performance evaluations can be discussed. This study performed multi-class classification. However, limited number of attributes used such as only genre attribute was used to describe a movie. Since a movie can be described with more than one genre, multi-class multi labels classifications can be adopted to classify multi-labels movies. Moreover, other movie attributes can be included in the classification to improve the study, which has not been included in this study because of real data usage. Due to the time limitation, to personalize each user's movie preferences, the change of the preferences over a specific time is not observed. As a future study, it is aimed to monitor user behavior by using a technique for example by looking to simple moving average for each user in specific time interval. The social and behavioral side of the recommendation system being developed may be interpreted. Specifically, the user acceptance of such system may be investigated. Last but not least, other real data sets can be used to assess the performance.

REFERENCES

Adomavicius, G., & Tuzhilin, A. (2005). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 734-749.

Ahn, H.J. (2008). A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. *Information Sciences*, *178 (1-2),* 37–51.

Aimeur, E., Brassard, G., Fernandez, J. M., & Onana F.S.M. (2006). Privacy-preserving demographic filtering. *in Proceedings of the ACM symposium on Applied computing, (*pp. 872–88*)*. New York, NY, USA: ACM.

Amatriain X., Jaimes A., Oliver N., & Puriol J.M. (2011). Data mining methods for recommender systems. In: Ricci F, Rokach L, Shapira B, Kantor P (Eds.), Recommender systems handbook (pp. 39–71). Berlin: Springer.

Berry, M. W., and Browne, M., (2006), (Eds.), *Lecture Notes in Data Mining* World Scientific, Singapore.

Bijak, K. & Thomas, L. C. (2012) Does segmentation always improve model performance in credit scoring?. *Expert Systems with Applications, 39*(3),2433-2442.

Billsus, D., & Pazzani, M. (1996). Revising User Profiles: The Search for Interesting Websites. *In Proceedings of 3rd International Workshop on Multistrategy Learning.*

Bobadilla, J., Ortega, F., Hernando, A., & Bernal, J. (2012). A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-Based Systems, 26*, 225-238.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). Classification and Regression Trees, Belmont, CA: Wadsworth International Group.

Burke, R. (2002). Hybrid recommender systems: survey and experiments. *User Modelling and User-Adapted Interaction*, *12 (4)*, 331–370.

Casali , A., Godo, L., Sierra, C. (2008). Validation and Experimentation of a Tourism Recommender Agent based on a Graded BDI Model.*Proceedings of the 11th International Conference of the Catalan Association for Artificial Intelligence, CCIA 2008 ,* (pp. 41-50).

Chen, E. (2011). Choosing a machine learning classifier. Retrieved March 04, 2013, from http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/.

Chen, Q. & Aickelin, U. (2004). *Movie Recommendation Systems Using an Artificial Immune System*. In Poster Proceedings of ACDM 2004 Engineers House, Bristol, UK.

Chen, T. & He, L. (2009). Collaborative Filtering Based on Demographic Attribute Vector. *ETP International Conference on Future Computer and Communication, (pp. 225-229).*

Chickering, D.M. (1996). Learning Bayesian Networks is NP-Complete. In V, D. Fisher and H. Lenz, (Eds.),*Learning from Data: Artificial Intelligence and Statistics*, (pp. 121-130). Springer-Verlag.

Chikhaoui B., Chiazzaro, M. & Wang, S.(2011). An Improved Hybrid Recommender System by Combining Predictions. *Proceedings of the 2011 IEEE Workshops of International Conference on Advanced Information Networking and Applications*, (pp.644-649).

Claypool, M., Gokhale, A., Miranda T., Murnikov, P., Netes, D., & Sartin, M. (1999). Combining content-based and collaborative filters in an online newspaper, *Proceeding of the ACM SIGIR'99 Workshop on Recommender Systems*.

Costa, P. (2005). *Bayesian Semantics for the semantic web*. Doctoral dissertation, George Mason University.

Cremonesi, P., Koren, Y., & Turrin, R. (2010). Performance of recommender algorithms on top-n recommendation tasks. *Proceedings of the fourth ACM conference on Recommender systems.* Barcelona, Spain.

De Pessemier, T., Deryckere, T., & Martens, L. (2010) Extending the Bayesian Classifier to a Context-Aware Recommender System for Mobile Devices. *In: Fifth International Conference on Internet an Web Applications and Services,* Barcelona, Spain

Du, H. (2010). *Data Mining Techniques and Applications an Introduction*: Cengage Learning EMEA.

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

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

Engelbert, B. ,Morisse K., Hamborg K. C., Evaluation and user acceptance issues of a Bayesian classifier based TV Recommendation System. CARS-2012, vol. 889.

Gayatri, N., Nickolas, S., & Reddy, A., V. (2010) Feature Selection Using Decision Tree Induction in Class level Metrics Dataset for Software Defect Predictions. *Proceedings of the World Congress of Engineering and Computer Science* WCECS 2010, Vol: I, San Fransisco, USA.

Ghazanfar, M. & Prügel-Bennet, A. (2010). An Improved Switching Hybrid Recommender System Using Naive Bayes Classifier and Collaborative Filtering.

Proceedings of the International MultiConference of Engineers and Computer ScientistsIMECS 2010, Vol: I. Hong Kong.

Hall, M. (2007). A decision tree-based attribute weighting filter for naive Bayes. *Knowledge-Based Systems*, *20,* 120-126.

Han, J., and Kamber, M. (2006). *Data Mining: Concepts and Techniques*., USA: Morgan Kaufmann Publishers Second Edition.

Herlocker, J. L., Konstan, J. A.,Terveen, L. G. & Riedl, J. T. (2004) Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems, 22(1)*, 5–53.
Herlocker, J.L., & Konstan, J.A., (2001). Content-independent task focused recommendation. *IEEE Internet Computing*, *5,* 40-47.

Jannach, D., Zanker, M., Felfernig, A., Friedrich, G. (2010). Recommender Systems An Introduction. Cambridge University Press.

Jiang, L., Cai, Z., Wang, D., & Zhang, H. (2012). Improving Tree augmented Naive Bayes for class probability estimation. *Knowledge-Based Systems*, *26,* 239-245.

Lavanya, D. & Usha R.K. (2011) Analysis of Feature Selection with Classification: Breast Cancer Datasets. *Indian Journal of Computer Science and Engineering*, 2(5).

Khatri, M. (2012). A Survey of Naïve Bayesian Algorithms for Similarity in Recommendation Systems. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2 (5).

Kohavi, R. (1996). Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 202–207). Portland, OR: AAAI Press.

Konstan, J. A. & Riedl, J (2012) Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22 (1-2), 101-123.

Kotsiantis, S.B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica, 31*, 249-268.

Kotsiantis, S., Zaharakis, I., Pintelas, P. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review* 26, 159–190

Krulwich, B. (1997). LifestyleFinder: Intelligent user profiling using large-scale demographic data. *Artificial Intelligence Magazine*, *18*, 37-45.

Kubat, M., Flotzinger, D., & Pfurtscheller, G. (1993). Discovering patterns in EEG-signals: Comparative Study of few methods. *In proceedings of the European Conference on Machine Learning, Vol:667. (pp.* 366-371). Vienna: LNCS.

Magidson, J., & Vermunt, J. K. (2005). An extension of the CHAID tree-based segmentation algorithm to multiple dependent variables. In C. Weihs & W. Gaul (Eds.), *Classification: The ubiquitous challenge* (pp. 240–247). Heidelberg: Springer.

Markellou, P., Mousourouli, I., Sirmakessis, S. & Tsakalidis, A. (2005). Personalized E-commerce Recommendations. *Proceedings of IEEE International Conference on e-Business Engineering.*

McLaughlin, M.R. & Herlocker, J.L. (2004). A collaborative filtering algoritm and evaluation metric that accurately model the user experience,*Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)* (pp. 329-336).ACM.

McCaffrey, J. (2012), Classification and Prediction Using Neural Networks. *MSDN Magazine*. Retrieved February 18, 2013, from http://msdn.microsoft.com/en-us/magazine/jj190808.aspx.

Melville, P., Mooney, R.J., & Nagarajan, R (2002). In Rina Dechter, Richard S. Sutton (Eds.): *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence*, (pp. 187-192). Edmonton, Alberta, Canada.

Montaner, M., Lopez, B., & de la Rosa, J.L. (2003). A taxonomy of recommender agents on the internet. *Artificial Intelligence Review,19(4)*, 285–330.

NUNES, M. A. S. N. & HU, R. (2012). Personality-based Recommender Systems: An Overview. *Proceedings of the sixth ACM conference on Recommender systems (RecSys '12).* (pp. 5-7) New York: ACM.

Oz, T. (2007). Data Mining Algoritması Naïve Bayes. Retireved March 04, 2013, from http://www.tameroz.com/data-mining-algoritmasi-naive-bayes/

Park, D. H., Kim, H.K., Choi, Y., & Kim J. K. (2012) A literature review and classification of recommender systems research. *Expert Systems with Applications, 39(11)*, 10059-10072.

Pazzani, M.J. (1999) A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review, 13*, 393-408.

Questier, F., Put, R., Coomans, D., Walczak, B., Vander Heyden, Y. (2005). The use of CART and multivariate regression trees for supervised and unsupervised feature selection, *Chemometrics and Intelligent Laboratory Systems*, 76 (1), 45-54.

Ratanamahatana, C. & Gunopulos, D. (2003). Feature selection for the naive Bayesian classifier using decision trees. *Applied Artificial Intelligence17(5–6)*, 475–48.

Resnick, P., Iakovou, N., Sushak, M., Bergstrom, P. & Riedl J. (1994). GroupLens an open architecture for collaborative filtering of netnews. *Computer supported cooperative work conference.*

Ricci, F., Rokach, L., Shapira, B. & Paul B. Kantor (Eds.). (2011). *Recommender Systems Handbook*. Springer.

Robles, V., Larranaga, P., Pena, J.M., Marban, O., Crespo, J., & Pérez, M.S. (2003). Collaborative filtering using interval estimation naive Bayes. *LNCS (LNAI), 2663*, 46–53.

Robles, V., Larranaga, P., Menasalvas, E., Perez, M.S.,& Herves, V. (2003). Improvement of naive Bayes collaborative filtering using interval estimation. *Web Intelligence, WI 2003. Proceedings. IEEE/WIC International Conference on* , (pp. 168- 174).

Russell, S.J. & Norvig, P. (2002). *Artificial Intelligence: A Modern Approach (2nd Edition).* Prentice Hall.

Said, A., Plumbaum, T., De Luca, E. W., Albayrak, S. (2011).*A Comparison of How Demographic Data Affects Recommendation*, Poster and Demo Session presented at *UMAP 2011.*

Sarwar, B., Karypis, G., Konstan, J. A., Riedl, J. (2000). Analysis of recommendation algorithms for e-commerce. *Proceedings of the ACM E-commerce, (pp.158-167).*

Sampat, M. P, Patel, A.C., Wang, Y., Gupta, S., Kan C-W, Bovik, A. C, Markey, M. K (2009) Indexes for three-class classification performance assessment–an empirical comparison. *IEEE Trans Inf Technol Biomed*, *13*, 300–312.

Sharma N. & Mukherjee, S. (2012). A Novel Multi-Classifier Layered Approach to Improve Minority Attack Detection in IDS. *Procedia Technology*, *6*, 913-921.

Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '02), (pp.253-260). New York, NY, USA: ACM.

Schröder,G., Thiele, M., Lehner, W. (2011). Setting Goals and Choosing Metrics for Recommender System Evaluations. *RecSys 2011Workshop on HumanDecision Making in Recommender Systems(Decisions@RecSys'11) and UserCentric Evaluation of Recommender Systems and Their Interfaces -2 (UCERSTI 2) affiliated with the 5thACM Conference on Recommender Systems,* Chicago, IL,USA.

Shih, D.H., Yen, D.C., Lin H.C. & Shih, M.H. (2011). An implementation and evaluation of recommender systems for traveling abroad. *Expert Systems with Applications, 38 (12),* 15344-15355.

Sharma, N. & Mukherjee, S. (2012). A Novel Multi-Classifier Layered Approach to Improve Minority Attack Detection in IDS. *Procedia Technology*, 6, 913-921.

Shardanand, U. & Maes, P. (1995) Social information filtering: algorithms for automating 'Word of Mouth'. *Conference Human Factors in Computing Systems*

Sobecki, J. (2006). Implementations of Web-based Recommender Systems Using Hybrid Methods. *International Journal of Computer Science & Applications:Technomathematics Research Foundation. 3 (3),* 52 – 64.

Sokolova, M. & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management: an International Journal*, 45 (4), 427-437.

SPSS. (1999). Answer Tree Algorithm Summary. SPSS White Paper, USA, s.2.

Stritt, M., Tso, K.H.L., Schmidt-Thieme, L. (2007). Attribute Aware Anonymous Recommender Systems. *Advances in Data Analysis Studies in Classification, Data Analysis, and Knowledge Organization,* 497-504.

Tan, P.N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining,* Addison-Wesley.

Tan, P.N., Steinbach, M. & Kumar, V., (2006), Introduction to Data Mining, Pearson International Edition.

Ting, S.L., Ip, W.H., Tsang, A. H. C. (2011). Is Naïve Bayes a Good Classifier for Document Classification?**.***International Journal of Software Engineering and Its Applications, 5 (3),* 37.

Ujjin, S. & Bentley, P.J. (2002) Learning user preferences using evolution. *In: 4th Asia-Pacific conference on simulated evolution and learning (SEAL'02)*, Singapore.

Vozalis, M., & Margaritis, K. (2004). Collaborative filtering enhanced by demographic correlation, *in AIAI Symposium on Professional Practice in AI, of the 18th World Computer Congress.*

Wang, X & Zhou, C. (2012). A collaborative filtering recommendation algorithm using user implicit demographic information. *Computer Science & Education (ICCSE), 2012 7th International Conference on* , (pp. 935-939).

Wang, K., & Tan, Y. (2011). A new collaborative filtering recommendation approach based on Naive Bayesian method. In Tan, Y.; Shi, Y.; Chai, Y.; and Wang, G., (Eds.), *Advances in Swarm Intelligence* (pp. 218-227). Berlin: Springer-Verlag.

Wei, K., Huang, J., & Fu, S. (2007). A Survey of E-Commerce Recommender Systems. *Service Systems and Service Management, 2007 International Conference on* , (pp 1-5).

Wen, Z. (2008). Recommendation System Based on Collaborative Filtering.

Wikipedia,. 2013. Precision and Recall. Retrieved March 04, 2013,from http://en.wikipedia.org/wiki/Precision_and_recall.

Zhhang, S., Zhang, C. & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence: An International Journal, 17(*5-6), 375-381.

APPENDICES

Appendix A

Confusion Matrices and Precision Recall results for first three genres

Table 12: Confusion matrix of only first genre matches

Actual Genres	Predicted Genres					
	Comedy	Drama	Horror	Romance	Action	Sci-fi
Comedy	5	2		1		1
Drama	2	9	3			
Horror			3			
Romance		3	1	5		2
Action			1	3	5	
Sci-fi	1		1		1	3

Table 13: Precision and recall metric for first genre

Genre	Precision	Recall
Comedy	0,63	0,56
Drama	0,64	0,64
Horror	0,33	1,00
Romance	0,56	0,45
Action	0,83	0,56
Science-fiction	0,50	0,50
Average	0,58	0,62

Table 14: Confusion matrix of only second genre matches

Actual Genres	Predicted Genres					
	Comedy	Drama	Horror	Romance	Action	Sci-fi
Comedy	4	1		2	1	2
Drama	1	6		1		2
Horror		4	3			2
Romance			1	7		1
Action	2				4	
Sci-fi	2	1	2			3

Table 15: Precision and recall metrics for second genre

Genre	Precision	Recall
Comedy	0,44	0,40
Drama	0,50	0,60
Horror	0,50	0,33
Romance	0,70	0,78
Action	0,80	0,67
Science-fiction	0,30	0,38
Average	0,54	0,53

Table 16:Confusion matrix of only third genre matches

Actual Genres	Predicted Genres					
	Comedy	Drama	Horror	Romance	Action	Sci-fi
Comedy	8	3	3	2	1	1
Drama		3	1	1	1	1
Horror		1	2			1
Romance	1			3		1
Action	1					
Sci-fi	1	3		2	2	4

Table 17: Precision and recall metrics for third genre

Genre	Precision	Recall
Comedy	0,73	0,44
Drama	0,30	0,43
Horror	0,33	0,50
Romance	0,38	0,75
Action	0,00	0,00
Science-fiction	0,50	0,33
Average	0,37	0,41

Appendix B

Visualization of distribution of demographic attributes of records in the validation data in terms of age, education and occupation

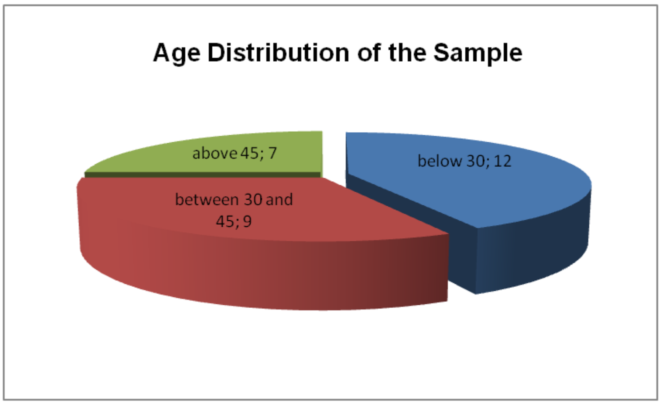


Figure 17: Age distribution of sample collected for validation

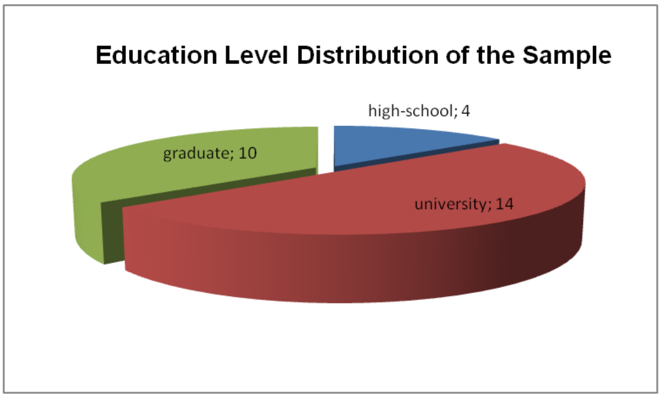


Figure 18: Education distribution of sample collected for validation

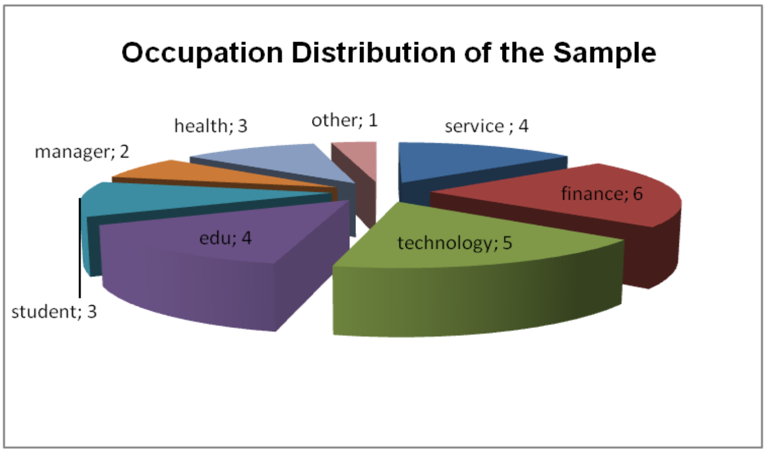


Figure 19: Occupation distribution of sample collected for validation

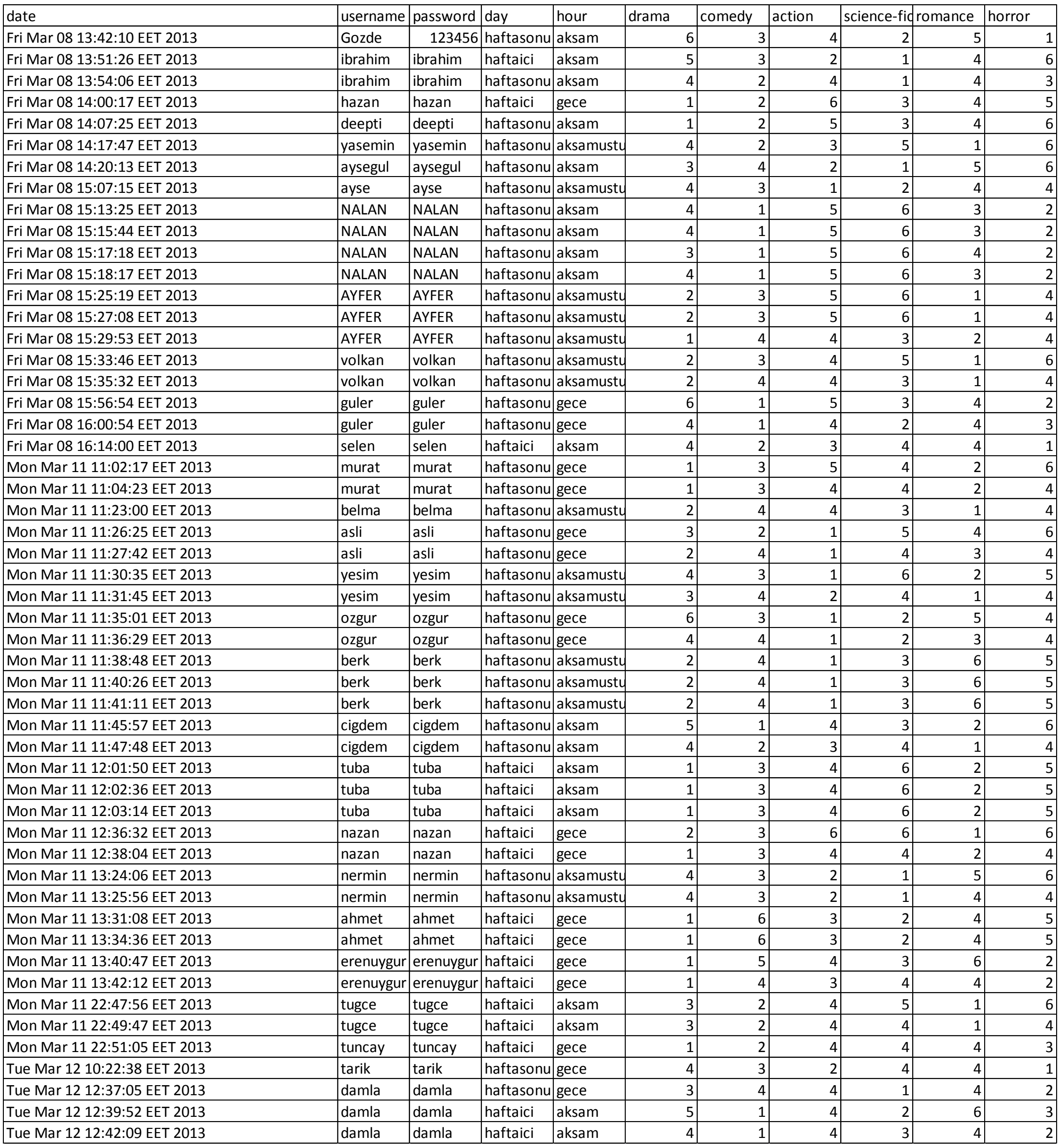
Appendix C

Data set for the validation

Table 18: User demographics

username	password	age	gender	education	occupation
erenuygur	erenuygur	below 30	male	graduate	other
asli	asli	below 30	female	graduate	finance
belma	belma	above 45	female	university	finance
berk	berk	below 30	male	university	finance
cigdem	cigdem	below 30	female	university	finance
ozgur	ozgur	between 30 and 45	male	university	finance
yesim	yesim	between 30 and 45	female	university	finance
ayse	ayse	between 30 and 45	female	highschool	service
aysegul	aysegul	between 30 and 45	female	highschool	service
murat	murat	above 45	male	university	service
yasemin	yasemin	between 30 and 45	female	university	service
hazan	hazan	below 30	female	graduate	student
tugce	tugce	below 30	female	university	student
tuba	tuba	below 30	female	graduate	student
damla	damla	below 30	female	graduate	education
deepti	deepti	between 30 and 45	female	graduate	education
guler	guler	between 30 and 45	female	graduate	education
ibrahim	ibrahim	_15ile30	male	graduate	education
AYFER	AYFER	above 45	female	graduate	health
NALAN	NALAN	above 45	female	university	health
volkan	volkan	below 30	male	highschool	health
ahmet	ahmet	above 45	male	university	technology
Gozde	123456	below 30	female	university	technology
selen	selen	between 30 and 45	female	university	technology
tarik	tarik	below 30	male	university	technology
tuncay	tuncay	above 45	male	graduate	technology
nazan	nazan	above 45	female	university	manager
nermin	nermin	between 30 and 45	female	highschool	manager

Table 19: User logs



date	username	password	day	hour	drama	comedy	action	science-fic	romance	horror
Fri Mar 08 13:42:10 EET 2013	Gozde	123456	haftasonu	aksam	6	3	4	2	5	1
Fri Mar 08 13:51:26 EET 2013	ibrahim	ibrahim	haftaici	aksam	5	3	2	1	4	6
Fri Mar 08 13:54:06 EET 2013	ibrahim	ibrahim	haftasonu	aksam	4	2	4	1	4	3
Fri Mar 08 14:00:17 EET 2013	hazan	hazan	haftaici	gece	1	2	6	3	4	5
Fri Mar 08 14:07:25 EET 2013	deepti	deepti	haftasonu	aksam	1	2	5	3	4	6
Fri Mar 08 14:17:47 EET 2013	yasemin	yasemin	haftasonu	aksamustu	4	2	3	5	1	6
Fri Mar 08 14:20:13 EET 2013	aysegul	aysegul	haftasonu	aksam	3	4	2	1	5	6
Fri Mar 08 15:07:15 EET 2013	ayse	ayse	haftasonu	aksamustu	4	3	1	2	4	4
Fri Mar 08 15:13:25 EET 2013	NALAN	NALAN	haftasonu	aksam	4	1	5	6	3	2
Fri Mar 08 15:15:44 EET 2013	NALAN	NALAN	haftasonu	aksam	4	1	5	6	3	2
Fri Mar 08 15:17:18 EET 2013	NALAN	NALAN	haftasonu	aksam	3	1	5	6	4	2
Fri Mar 08 15:18:17 EET 2013	NALAN	NALAN	haftasonu	aksam	4	1	5	6	3	2
Fri Mar 08 15:25:19 EET 2013	AYFER	AYFER	haftasonu	aksamustu	2	3	5	6	1	4
Fri Mar 08 15:27:08 EET 2013	AYFER	AYFER	haftasonu	aksamustu	2	3	5	6	1	4
Fri Mar 08 15:29:53 EET 2013	AYFER	AYFER	haftasonu	aksamustu	1	4	4	3	2	4
Fri Mar 08 15:33:46 EET 2013	volkan	volkan	haftasonu	aksamustu	2	3	4	5	1	6
Fri Mar 08 15:35:32 EET 2013	volkan	volkan	haftasonu	aksamustu	2	4	4	3	1	4
Fri Mar 08 15:56:54 EET 2013	guler	guler	haftasonu	gece	6	1	5	3	4	2
Fri Mar 08 16:00:54 EET 2013	guler	guler	haftasonu	gece	4	1	4	2	4	3
Fri Mar 08 16:14:00 EET 2013	selen	selen	haftaici	aksam	4	2	3	4	4	1
Mon Mar 11 11:02:17 EET 2013	murat	murat	haftasonu	gece	1	3	5	4	2	6
Mon Mar 11 11:04:23 EET 2013	murat	murat	haftasonu	gece	1	3	4	4	2	4
Mon Mar 11 11:23:00 EET 2013	belma	belma	haftasonu	aksamustu	2	4	4	3	1	4
Mon Mar 11 11:26:25 EET 2013	asli	asli	haftasonu	gece	3	2	1	5	4	6
Mon Mar 11 11:27:42 EET 2013	asli	asli	haftasonu	gece	2	4	1	4	3	4
Mon Mar 11 11:30:35 EET 2013	yesim	yesim	haftasonu	aksamustu	4	3	1	6	2	5
Mon Mar 11 11:31:45 EET 2013	yesim	yesim	haftasonu	aksamustu	3	4	2	4	1	4
Mon Mar 11 11:35:01 EET 2013	ozgur	ozgur	haftasonu	gece	6	3	1	2	5	4
Mon Mar 11 11:36:29 EET 2013	ozgur	ozgur	haftasonu	gece	4	4	1	2	3	4
Mon Mar 11 11:38:48 EET 2013	berk	berk	haftasonu	aksamustu	2	4	1	3	6	5
Mon Mar 11 11:40:26 EET 2013	berk	berk	haftasonu	aksamustu	2	4	1	3	6	5
Mon Mar 11 11:41:11 EET 2013	berk	berk	haftasonu	aksamustu	2	4	1	3	6	5
Mon Mar 11 11:45:57 EET 2013	cigdem	cigdem	haftasonu	aksam	5	1	4	3	2	6
Mon Mar 11 11:47:48 EET 2013	cigdem	cigdem	haftasonu	aksam	4	2	3	4	1	4
Mon Mar 11 12:01:50 EET 2013	tuba	tuba	haftaici	aksam	1	3	4	6	2	5
Mon Mar 11 12:02:36 EET 2013	tuba	tuba	haftaici	aksam	1	3	4	6	2	5
Mon Mar 11 12:03:14 EET 2013	tuba	tuba	haftaici	aksam	1	3	4	6	2	5
Mon Mar 11 12:36:32 EET 2013	nazan	nazan	haftaici	gece	2	3	6	6	1	6
Mon Mar 11 12:38:04 EET 2013	nazan	nazan	haftaici	gece	1	3	4	4	2	4
Mon Mar 11 13:24:06 EET 2013	nermin	nermin	haftasonu	aksamustu	4	3	2	1	5	6
Mon Mar 11 13:25:56 EET 2013	nermin	nermin	haftasonu	aksamustu	4	3	2	1	4	4
Mon Mar 11 13:31:08 EET 2013	ahmet	ahmet	haftaici	gece	1	6	3	2	4	5
Mon Mar 11 13:34:36 EET 2013	ahmet	ahmet	haftaici	gece	1	6	3	2	4	5
Mon Mar 11 13:40:47 EET 2013	erenuygur	erenuygur	haftaici	gece	1	5	4	3	6	2
Mon Mar 11 13:42:12 EET 2013	erenuygur	erenuygur	haftaici	gece	1	4	3	4	4	2
Mon Mar 11 22:47:56 EET 2013	tugce	tugce	haftaici	aksam	3	2	4	5	1	6
Mon Mar 11 22:49:47 EET 2013	tugce	tugce	haftaici	aksam	3	2	4	4	1	4
Mon Mar 11 22:51:05 EET 2013	tuncay	tuncay	haftaici	gece	1	2	4	4	4	3
Tue Mar 12 10:22:38 EET 2013	tarik	tarik	haftasonu	gece	4	3	2	4	4	1
Tue Mar 12 12:37:05 EET 2013	damla	damla	haftasonu	gece	3	4	4	1	4	2
Tue Mar 12 12:39:52 EET 2013	damla	damla	haftaici	aksam	5	1	4	2	6	3
Tue Mar 12 12:42:09 EET 2013	damla	damla	haftaici	aksam	4	1	4	3	4	2

Appendix D

QUESTIONNAIRE

Kullanıcının Film Türü Seçimini belirleme anketi

1. Film seçerken aşağıdakilerden hangisini/ hangilerini dikkate alırsınız? Birden çok maddeyi seçebilirsiniz.
   - □ Gün
     - □ Haftaiçi
     - □ Haftasonu
   - □ Saat
     - □ 11:00-15:00
     - □ 16:00-17:00
     - □ 18:00-19:00
     - □ 20:00-24:00
   - □ Yer
     - □ Uzaklık
     - □ Sinema salonu tipi

2. Aşağıda verilen maddeleri film seçerken, en çok önemsediğinizden en az önemsediğinize göre numaralandırınız.

   □ Film Türü (dram, komedi, aksiyon, bilimkurgu, romantik,korku vb.)
   □ Dili (Yerli, yabancı)
   □ Oyuncu kadrosu (Aktör/ Aktris)
   □ Yönetmen
   □ Ödül (Oscars, Altın Portakal, vb.)
   □ Yapım yılı

3. Film türü tercihiniz kiminle film izlediğinize göre değişir mi?

   □ Evet (Kişi: ...., Tür:........................)
   □ Hayır

Katılımınız için teşekkür ederiz ☺

Appendix E

In following figures 20, 21, 22, 23 the demographic distribution of the original purchase data's record sample are presented.

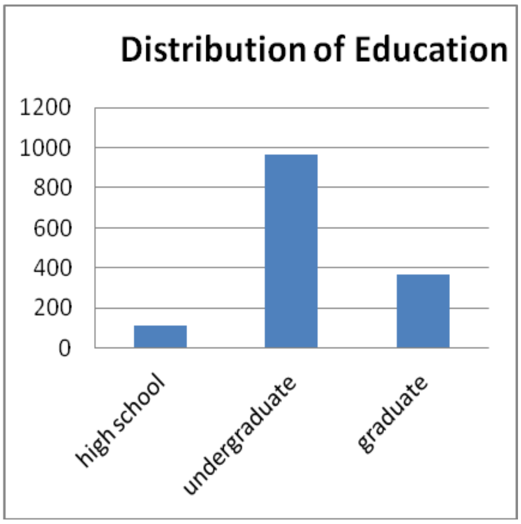


Figure 20: Education Distribution of the sample

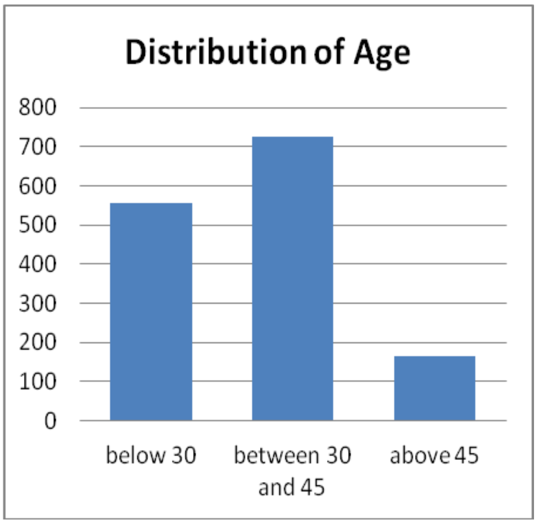


Figure 21: Age Distribution of the sample

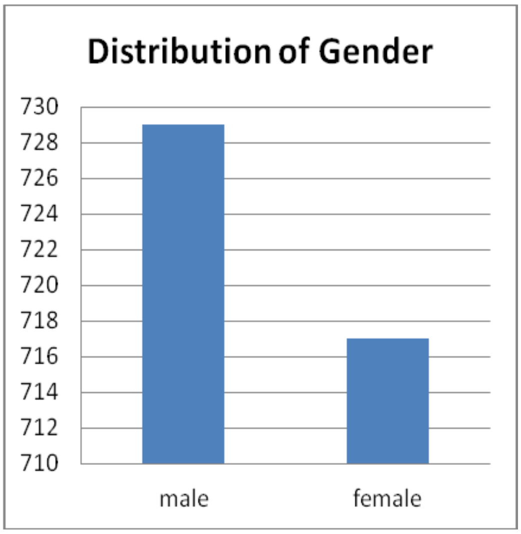


Figure 22: Gender Distribution of the sample

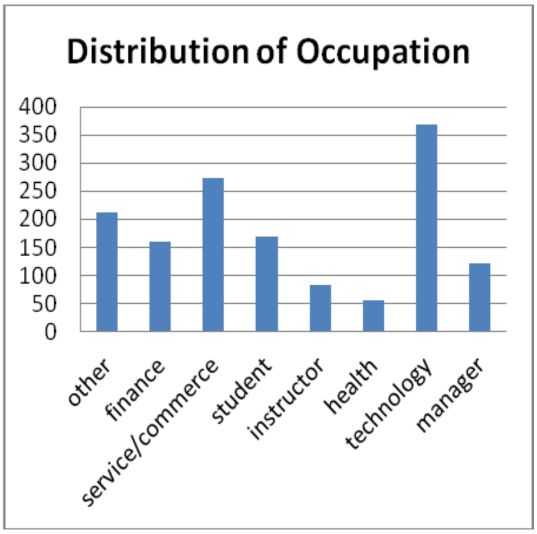


Figure 23: Occupation Distribution of the sample

According to the Figures 20 and 21, it can be said that different sub-categories of the attributes such as education and age are not equally distributed.

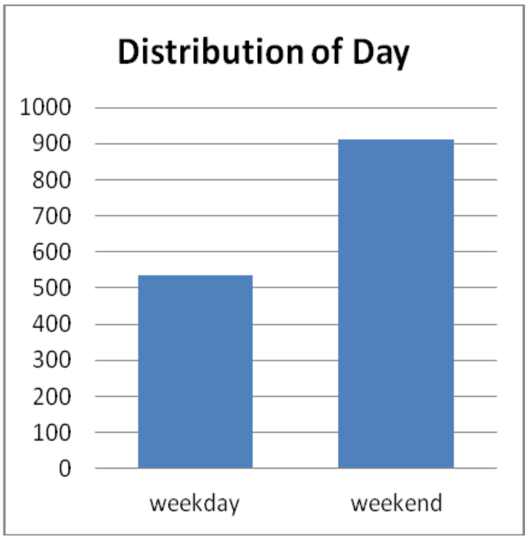


Figure 24: Day Distribution of the sample

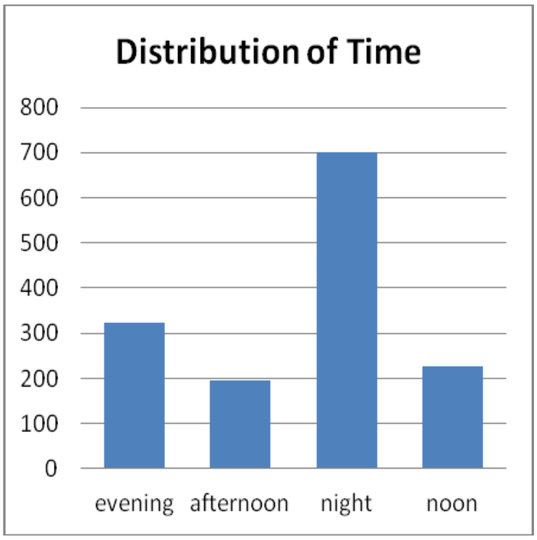


Figure 25: Time Distribution of the sample

Figure 24 and 25 show the day and time distributions of people booked a movie ticket. From the figures it can be concluded that people naturally prefer to go the cinema during weekends. When we look to the time slots, it can be seen easily that people prefer night hours for watching a movie. These facts can be explained with the reason that most of the people are working during the weekdays and day hours and therefore they prefer weekends and night hours mostly.

Appendix F

Attributes		Movie Genres					
		Sci-fi	drama	horror	comedy	action	romance
day	weekday	72	101	99	85	88	90
	weekend	169	140	142	156	153	151
hour	10:00-15:00	43	39	31	44	33	37
	16:00-17:00	46	36	16	27	31	39
	18:00-19:00	49	47	57	62	52	57
	20:00-24:00	103	119	137	108	125	108
occupation	Other	32	44	42	27	31	36
	Finance	22	35	26	20	22	36
	service-commerce	44	37	39	47	55	52
	student	29	26	28	27	30	29
	instructor	14	17	13	16	10	13
	health	13	12	10	8	7	7
	technology	65	49	70	68	65	51
	manager	22	21	13	28	21	17
education	high school	15	21	21	22	22	14
	undergraduate	158	153	159	160	160	176
	graduate	68	67	61	59	59	51
gender	Male	118	125	119	126	126	115
	female	123	116	122	115	115	126
age	below 30	88	78	108	86	94	102
	30 - 45	127	119	118	120	132	110
	above 45	26	44	15	35	15	29

Appendix G

Simple demonstration for the formula

$$Weight\ \ of\ \ Genre\ =P(genre|X)+(P(genre|X)*\frac{6}{avg(genre\ \ rank)})$$

Consider a user X with following classification results according to NBC for each movie genres:

Table 20: Genre weight calculation for recommendation

NBC	Average Ranks	Weights of genres
P (romance) =0.50	4	0.50+0.50*6/4=1.25
P(sci-fi) = 0.45	5	0.45+0.45*6/5=0.99
P (action) =0.40	2	0.40+0.40*6/2=1.6
P (drama) =0.20	1	0.20+0.20*6/1=1.40
P(horror)=0.23	6	0.23+0.23*6/6=0,46
P (comedy)=0.10	3	0.10+0.10*6/3=0,30

Orders of the movies are as follows the orders did not changed very drastically, since the weight is dependent on both the posterior probability of the genre and the rank the specific user assigns to it.

Table 21: Genre orders according to NBC, ranks and weights

According to NBC	According to ranks	According to formula
Romance	Drama	Action
Sci-fi	Action	Drama
Action	Comedy	Romance
Drama	Romance	Sci-fi
Horror	Sci-fi	Horror
Comedy	Horror	Comedy

TEZ FOTOKOPİSİ İZİN FORMU

ENSTİTÜ

Fen Bilimleri Enstitüsü □

Sosyal Bilimler Enstitüsü □

Uygulamalı Matematik Enstitüsü □

Enformatik Enstitüsü ☒

Deniz Bilimleri Enstitüsü □

YAZARIN

Soyadı : HACALOĞLU
Adı : TUNA
Bölümü : Bilişim Sistemleri

TEZİN ADI (İngilizce) : INTERNET BASED MOVIE GENRE SUGGESTION MODEL CONSIDERING DEMOGRAPHICAL INFORMATION OF USERS

TEZİN TÜRÜ : Yüksek Lisans ☒ Doktora □

1. Tezimin tamamından kaynak gösterilmek şartıyla fotokopi alınabilir. □
2. Tezimin içindekiler sayfası, özet, indeks sayfalarından ve/veya bir bölümünden kaynak gösterilmek şartıyla fotokopi alınabilir. ☒
3. Tezimden bir (1) yıl süreyle fotokopi alınamaz. □

TEZİN KÜTÜPHANEYE TESLİM TARİHİ :