http://www.calsec.or.kr/jsebs
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.19, No.3, Aug 2014, pp.125-141
http://dx.doi.org/10.7838/jsebs.2014.19.3.125

IMDB 사용자평점에 대한 인구통계학적 분석의 활용

Utilization of Demographic Analysis with IMDB User Ratings on the Recommendation of Movies

배성문(Sung Moon Bae)*, 이상천(Sang Chun Lee)**, 박종훈(Jong Hun Park)***

초 록

인터넷에서 매 순간 발생하는 데이터의 홍수는 사용자가 필요로 하는 유용한 정보를 검색하는데 어려움을 초래한다. 그래서 많은 사용자들이 자신이 원하는 정보를 쉽게 찾기 위한 기법을 고안하고 이를 지원하는 도구를 개발하게 되었다.

이런 유용한 도구 중 하나인 추천시스템은 기존의 사용자 정보를 분석하여 사용자가 원하는 제품이나 정보를 추천하는 것이다. 본 논문에서는 추천시스템을 활용하여 원하는 정보를 제안하는데 인구통계학적인 기법을 사용한다. 인구통계학 기반 추천시스템은 나이, 성별과 같은 인구통계학적인 특성을 사용하여 유용한 정보를 추출한다. 본 연구는 영화 선택 시 중요한 요소인 사용자 평점을 분석하고 이를 활용할 수 있는 방법을 제시하였다. 이를 위해 Internet Movie Database(IMDB) 웹 사이트에 있는 영화의 사용자 평점을 인구통계학적 요인으로 분석하였다.

본 논문에서는 인구통계학적 분석을 위해 사용자를 성별과 연령대로 분류하였고, 각 영화를 22개 장르로 나눈 IMDB 기준에 따라 사용자 평점을 분석하였다. 각 장르별 영화에 대해 사용자 그룹의 평균 평점을 F-테스트와 T-테스트를 수행하여 그 장르 영화 평점과 동일한 결과를 나타내는 대표 그룹을 찾아내었다. 인구통계학적 분석 결과인 대표 그룹은 새로운 영화가 개봉될 때 대표 그룹에 대한 프로모션과 추천을 통해 영화 홍보를 할 수 있는 대상을 찾아내는데 유용하다.

ABSTRACT

Nowadays, overflowing data produced every second from the internet make people to be difficult to search for the useful information. That's why people have invented and developed unique tools that they get some relevant information.

In this paper, the recommender system, one of the effective tools, is used and it helps us to get the useful information that we want by using demographic information to predict new items of interest. The demographic recommender system in this paper computes users' similarity using demographic information, age and gender. So we performed demographic analysis on movie ratings on Internet Movie Database (IMDB) web site that movies are rated by thousands of people, where users submitted a movie rating after they watched

* Department of Industrial and Systems Engineering/Engineering Research Institute, Gyeongsang National University(bsm@gnu.ac.kr)

** Corresponding Author, Department of Industrial and Systems Engineering/Engineering Research Institute, Gyeongsang National University(sclee@gnu.ac.kr)

*** Department of Business Administration, Catholic University of Daegu(icelatte@cu.ac.kr)

2014년 07월 03일 접수, 2014년 08월 22일 심사완료 후 2014년 08월 27일 게재확정.

a recent popular film. Meanwhile, we can understand that user's ratings, among various determinants of box office, is very essential factor in the study on recommendation of movie. This paper is aimed at analyzing movie average ratings directly given by film viewers, categorizing them into groups by sex and age, investigating the entire group and finding the representative group by examining it with F-test and T-test. This result is used to promote and recommend for the target group only.

Therefore, this study is considerably significant as presenting utilization for movie business as well as showing how to analyze demographic information on movie ratings on the web.

키워드 : 추천시스템, 인구통계학적 분석, 사용자 평점, IMDB
Recommender System, Demographic Analysis, User Ratings, IMDB

1. Introduction

The wide spreading of internet use brought the rapid development of e-business, which both companies and customers is required the change of pattern [16]. In addition, the rapid growth of Internet has produced huge amounts of data that the users cannot manage directly unless they use some tools. To find the information the users want among these numerous data, they spend a lot of time and endeavor. That's why people have invented and developed unique tools that they get some relevant information. Lee showed the impact of eWOM on consumer decision making process by viewing eWOM as the product information supplier [15]. Web users cannot live without well-optimized search engines, as we constantly find ourselves in the presence of the intelligent web and its main engine, the recommender system [18]. Recommender systems are powerful tools which suggest useful information that an Internet user may be interested in. Most recommender systems use collaborative filtering or content-based methods or demographic information to predict new items of interest for a user. Demographic recommender systems compute users' similarity using demographic information (age, location, profession, education, etc.) [20].

The significance of the image content industry, as a flagship industry of the cultural content industry, is emerging globally as a high value-added trend. Since the movie industry has diverse related industries along with a positive spillover effect from cultural and economic perspectives, it ultimately strengthens national brands as well as national competitiveness [3]. The movie industry has been sharply developing as a high value-added industry, and the numbers of movie viewers are also on the rise every year. According to the statistic data of Korean film council in 2012 about the movie industry, total viewers is about 194,890 thousand people, 1.9% up compared with last year and sales is about 1.455 trillion won, 17.7% up compared with last year [3]. Also, the age groups of movie viewers have come to vary. Since the

movie industry creates infinite demands and has remarkable potential for further development, movie distributors make a lot of resourceful investment for a hit. They perform promotion and marketing strategies in various media even before opening of a movie. The factors that affect a success of a movie are production cost, the number of running screens, prize winning at film festivals, ratings of professional film critics, ratings of netizens, word-of-mouth, directors, main characters, and genres.

Audiences do not know fully what pleasures they are going to get before they actually experience a film [7]. Therefore, the satisfaction, reviews, and average ratings by movie viewers are checked to evaluate movies, and the evaluated data are based in terms of movie choice. Prior to introduction of the internet, those who had information in every area held power so that a few controlled the distribution and contents of information. Also, professional movie critics and journalists, as early adapters, recommended movies and suggested the ways to watch movies through the typical press and broadcasting media [4].

Film critics are regarded as a significant factor to the box office performance. In these days, however, online reviews written by ordinary audiences are viewed as another important factor to the box office performance. They also make personal judgments with positive or negative evaluations on the film. As online reviews become meaningful either for the audience or for the box office result, we can sometimes witness critical difference of evaluation between audiences and critics [19].

This work is aimed at analyzing the data about movie average ratings directly given by netizens, categorizing them into groups by sex and age, finding and investigating the entire group and the most similar group by genre, and proposing that movie promotion, marketing, and recommendation strategies should be performed on the group.

This paper is organized as follows: Section 2 describes related works-recommender systems, collaborative filtering, content-based technique, and demographic filtering. Section 3 shows the user ratings system in IMDB and demographic information provided in IMDB. Section 4 presents the result of demographic analysis of user ratings. Section 5 describes the relationships demographic information and genres of movie and a recommendation of movies. Finally conclusions are presented and some open issues are discussed in Section 6.

2. Related Work

2.1 Recommender Systems

Recommender systems are programs which attempt to predict items that a user may be interested in, given some information about the user's profile [28]. Recommender systems

suggest personalized recommendations on items to users based on various kinds of information on users and items. Recommender systems intend to model user preferences on items and aim to recommend such items the user will probably like. User preference information is twofold : explicit and implicit feedback. The former mainly includes opinion expression via ratings of items on a predefined scale, while the latter consists of other user activities, such as purchasing, viewing, renting or searching of items [21].

The purpose of many studies on recommender systems have mainly focused on their capability of how likely they are able to recommend products that a customer is satisfied with. The first recommender systems appeared during the early 90's, mostly based on and expanding the terminology of collaborative filtering [13]. Later, when both the numbers of researches and the potential usage areas had grown, the scope of the terminology broadened. Four fundamental approaches to recommendation can be mentioned : demographic filtering, collaborative and content-based recommendation, and simplified statistical approaches [10]. However, recommender systems now use the following five technologies [18] :

- Collaborative Filtering Recommender Systems : as there is a given user, find another user or user group who has similar behavior to predict items of interest.
- Content-based Recommender Systems : technology usually employs a classifier to predict items' similarity; Recommend an item or a service to a user based upon the properties of the item and a profile of the user's interests.
- Demographic Recommender Systems : compute users' similarity using demographic information (age, location, profession, education, etc.).
- Knowledge-based Recommender Systems : build a knowledge base with a model of the users and/or items in order to apply inference techniques and find matches between users' need and items' features.
- Utility based Recommender Systems : Utility-based recommenders make suggestions based on a computation of the utility of each object for the user.

Several researches mentioned others category of systems; Reference [2] introduces temporal recommender systems, which are intended for suggesting items in situations where time is an essential factor of the decision-making process. It will be essential for the application in dynamic mobile recommender systems.

2.2 Content-based Recommender Systems

Content-based recommendation focuses on the similarity between products, usually taking into account their features like textual de-

scriptions [9], hyperlinks, related ratings [17], or co-occurrence in the same purchased transactions or web user sessions [8]. Content-based methods make recommendations by analyzing the description of the items that have been rated by the user and the description of items to be recommended. A variety of algorithms have been proposed for analyzing the content of text documents and finding regularities in this content that can serve as the basis for making recommendations. Many approaches are specialized versions of classification learners, in which the goal is to learn a function that predicts which class a document belongs to (i.e., either liked or not-liked). Other algorithms would treat this as a regression problem in which the goal is to learn a function that predicts a numeric value (i.e., the rating of the document). There are two important sub problems in designing a content-based filtering system. The first is finding a representation of documents. The second is to create a profile that allows for unseen documents to be recommended [20].

2.3 Collaborative Filtering Recommender Systems

Collaborative recommendation is typically based on item ratings explicitly delivered by users. The system recommends products, which have been evaluated positively by another similar user or by a set of such users, whose ratings have the strongest correlation with the current user [6]. Collaborative filtering (CF) [12, 22] is the most successful recommender system technology to date, and is used in many of the most successful recommender systems on the Web. CF systems recommend products to a target customer based on the opinions of other customers. These systems employ statistical techniques to find a set of customers known 0neighbors, that have a history of agreeing with the target user (i.e., they either rate different products similarly or they tend to buy similar set of products). Once a neighborhood of users is formed, these systems use several algorithms to produce recommendations [24].

Some hybrid recommender systems attempt to combine the advantages of two or more of the techniques [1, 25, 27]. Reference [1] presents a hybrid recommender system using a new heuristic similarity measure for collaborative filtering that focuses on improving performance under cold-start conditions where only a small number of ratings are available for similarity calculation for each user.

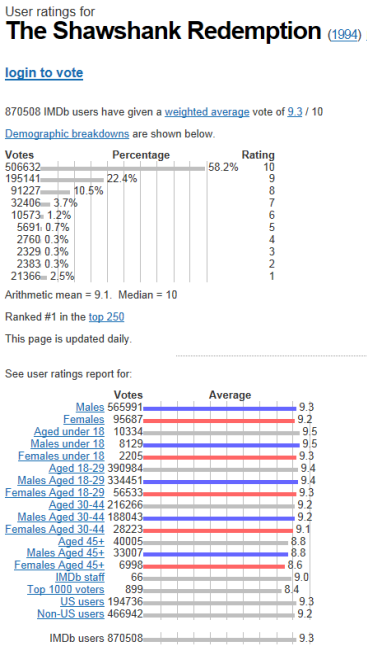
2.4 Demographic Recommender Systems

A user's profile consists simply of the data that the user has specified. These data are compared to those of other users to find overlaps in interests between users, and each user is recommended new items from the data of other users with overlapping interests. This

approach requires less computation than the previous one because it doesn't have to reason about the user data, and it clearly leverages the commonalities between users. However, it has the drawbacks of requiring data from a large number of users before being effective, requiring a large amount of data from each user, and limiting its recommendations to the exact items specified by the population of users [14]. An early example of this kind of system was Grundy that recommended books based on personal information gathered through an interactive dialogue [23]. Users' responses were matched against a library of manually assembled user stereotypes. Some more recent recommender systems have also taken this approach. Krulwich, for example, uses demographic groups from marketing research to suggest a range of products and services [14]. A short survey is used to gather the data for user categorization. In other systems, machine learning is used to arrive at a classifier based on demographic data [20]. Reference [26] shows how Singular Value Decomposition (SVD) along with demographic information can enhance plain Collaborative Filtering algorithms. Alternatively, this data can be extracted from the purchasing history, survey responses, etc. Each product is assigned to one or more classes with certain weights and the user is attracted to items from the class closest to their profile. This is attribute based recommendation [11]. In this study, demographic recommendation techniques using IMDB data are used. Recommended the group classified by the target group as age group, gender group.

3. IMDB User Ratings and User Profiles

Data used in this study was extracted from the world largest Internet Movie Database. The IMDB is an internet movie database owned by Amazon.com and it is an on-line database of various fields including movies, actors, TV soap operas, video games. Users submitted user ratings and reviews after they watched a recent popular film. The user reviews were used to increase customer satisfaction in [5].


User ratings for
The Shawshank Redemption (1994)

login to vote

870508 IMDb users have given a weighted average vote of 9.3 / 10

Demographic breakdowns are shown below.

Votes	Percentage	Rating
506632	58.2%	10
195141	22.4%	9
91227	10.5%	8
32406	3.7%	7
10573	1.2%	6
5691	0.7%	5
2760	0.3%	4
2329	0.3%	3
2383	0.3%	2
21366	2.5%	1

Arithmetic mean = 9.1. Median = 10

Ranked #1 in the top 250

This page is updated daily.

See user ratings report for:

	Votes	Average
Males	565991	9.3
Females	95687	9.2
Aged under 18	10334	9.5
Males under 18	8129	9.5
Females under 18	2205	9.3
Aged 18-29	390984	9.4
Males Aged 18-29	334451	9.4
Females Aged 18-29	56533	9.3
Aged 30-44	216266	9.2
Males Aged 30-44	188043	9.2
Females Aged 30-44	28223	9.1
Aged 45+	40005	8.8
Males Aged 45+	33007	8.8
Females Aged 45+	6998	8.6
IMDb staff	66	9.0
Top 1000 voters	899	8.4
US users	194736	9.3
Non-US users	466942	9.2
IMDb users	870508	9.3

〈Figure 1〉 IMDB User Ratings Report

<Figure 1> shows an example of user ratings report of a movie. Ratings of top 50 movies for each of 22 genres were analyzed among above database and the subject of this study is data on IMDB web site as of Nov. 2011. Data collected is the rating of top 50 movies for each genre and the ratings classified into each category. Average number of participants who gave ratings for each genre is 4,429,239 and males aged 18~29 is a group which gave all ratings in 22 genres.

Ratings are given from 1~10 point and types of genre include Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, Film-Noir, History, Horror, Music, Musical, Mystery, Romance, Sci-Fi, Sport, Thriller, War, and Western. There is rating data for each category and each genre and the category of subject group which gives classified rating is as <Table 1>. In here, insignificant categories including IMDB staff, Top 1000 voters, US users, and Non-US users group are excluded.

〈Table 1〉 Category of Subject Group which Gives Rating

Gender	Age Range	Gender-Age Range Pairs
Males/ Females	Aged under 18	Males under 18
		Females under 18
	Aged 18~29	Males Aged 18~29
		Females Aged 18~29
	Aged 30~44	Males Aged 30~44
		Females Aged 30~44
	Aged 45+	Males Aged 45+
		Females Aged 45+

There are total ratings for each movie and the category is classified into male and female based on gender. Classifications based on age include aged under 18, aged 18~29, aged 30~44, and aged 45+. Classifications based on age are classified once again based on gender for each age group thus the ratings are given by total 8 groups classified based on age and gender.

4. Analysis of Demographic Data

4.1 Analysis on Basic Statistics

In regards to top 50 movies of each genre, minimum value, second quartile value, median, fourth quartile value, mean, maximum value, variance, standard deviation, skewness of each category were extracted. <Table 2> illustrates total ratings and standard deviation of each genre.

Looking into total ratings data of each genre, average ratings are 6.2~7.8 and skew value of all genres except 4 was presented to be negative. skewness presents the degree of asymmetry of the distribution and it is calculated based on following formula.

$$S_k = \sum_{i=1}^{n} \frac{[(x_i - x_m)/s]^3}{n-1} \qquad (1)$$

where, x_m : sample mean,

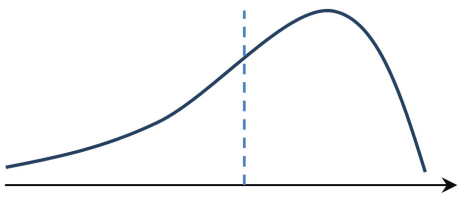
s : standard deviation of the sample.

〈Table 2〉 Average User Ratings and Skewness of Each Genre

genre	Avg. user ratings	Stdev.	Skew-ness
Action	7.4	0.9	0.2
Adventure	7.5	0.9	-0.6
Animation	7.2	0.9	-1.1
Biography	7.4	0.7	-0.1
Comedy	6.9	0.9	-0.4
Crime	7.2	1.0	-0.3
Documentary	7.7	1.8	-1.4
Drama	7.5	1.1	-0.8
Family	7.0	1.4	-1.1
Fantasy	6.9	1.4	-0.2
Film-Noir	7.8	0.5	-0.8
History	7.3	0.7	-0.4
Horror	6.2	0.8	0.0
Music	6.3	1.2	-0.9
Musical	6.8	1.0	-1.2
Mystery	7.1	0.8	-0.3
Romance	6.9	0.9	-0.2
Sci-Fi	7.1	1.0	0.3
Sport	6.6	0.9	-0.3
Thriller	6.9	0.9	0.2
War	7.1	1.0	-1.6
Western	6.6	0.9	-0.2

Form of distribution is as following based on skewness value.

- s = 0 : Normal Distribution
- s < 0 : Negative Skew with tail toward Left
- s > 0 : Positive Skew with Tail toward Right



〈Figure 2〉 Shape of Negative Skewness Value

As skewness of 18 genres was negative thus the majority is the negative skew with tail toward left. <Figure 2> shows the shape of the negative skewness value.

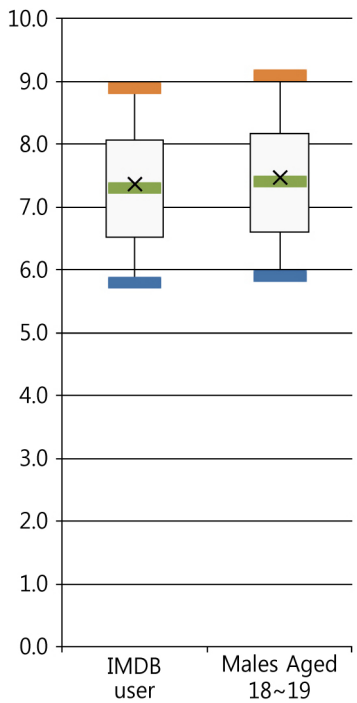
4.2 Boxplot of Two Similar Groups

Distribution of rating data and group with most similar distribution were found by creating boxplot with minimum value, second quartile value, median, fourth quartile, mean, and maximum value acquired through analysis on basic data. Then, a group with most similar value with total ratings in regards to mean, median, standard deviation, and skew value was found. <Table 3> illustrates the value of males aged 18~29 which has most similar value in mean, median, standard deviation, and skew in Action genre. As a result of comparing basic statistics of two groups, it was presented that they show difference within range of ±0.1. As you can see from Picture 1 illustrating the boxplot of two groups (All users, Males Aged 18~29), the distribution of sample for total ratings and Males Aged 18~29 group is very similar visually.

〈Table 3〉 Comparison on Basic Statistics with Average User Ratings (Action)

	All users	Males Aged 18~29
Mean	6.6	6.6
Median	6.6	6.7
Variance	0.8	0.9
Stdev.	0.9	0.9
Skewness	-0.2	-0.1

<Figure 3> shows a similarity of all users group and males aged 18~29 group. Two groups have very similar values of mean, median, variance, standard deviation, and skewness. <Table 4> illustrates the group most similar with total ratings acquired through analysis on basic statistics of each of 22 genres. As a result of extracting group most similar with total ratings for each genre using basic statistics and boxplot, Males Aged 18~29 group was presented to be most similar group in 13 out of 22 genres. This means that influence of Aged 18~29 group is most significant among groups which give ratings.



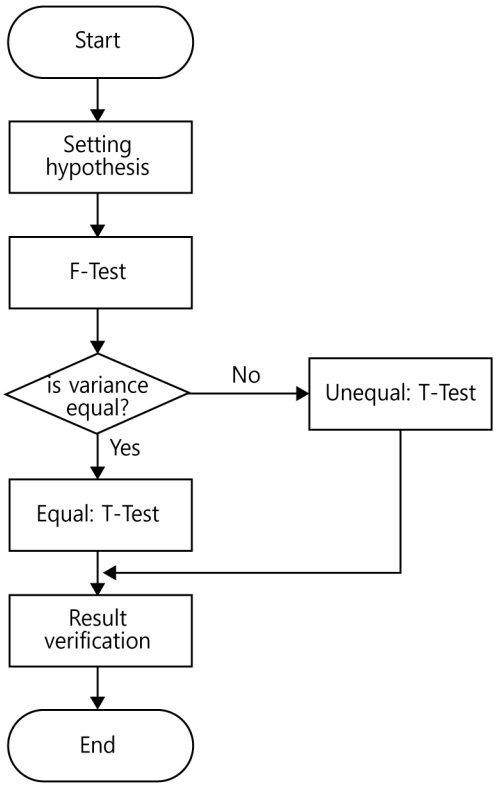
〈Figure 3〉 Boxplot of Two Similar Groups

〈Table 4〉 Representative Group of Each Genre

genre	User ratings	representative group
Action	7.4	Males Aged 18~29
Adventure	7.5	Males Aged 18~29
Animation	7.2	Males Aged 45+
Biography	7.4	Males Aged 18~29
Comedy	6.9	Males Aged 30~44
Crime	7.2	Males Aged 18~29
Documentary	7.7	Males Aged 30~44
Drama	7.5	Males Aged 18~29
Family	7.0	Males Aged 18~29
Fantasy	6.9	Males Aged 18~29
Film-Noir	7.8	Males Aged 18~29
History	7.3	Females Aged 30~44
Horror	6.2	Males Aged 18~29
Music	6.3	Males Aged 18~29
Musical	6.8	Males Aged 45+
Mystery	7.1	Males Aged 18~29
Romance	6.9	Males Aged 45+
Sci-Fi	7.1	Males Aged 18~29
Sport	6.6	Females Aged 30~44
Thriller	6.9	Males Aged 30~44
War	7.1	Males Aged 18~29
Western	6.6	Males Aged 18~29

4.3 F-Test and T-Test

In order to test whether or not a group most similar with total ratings based on basic statistics can represent total group, T-test which examines the significant difference between the mean of two groups was conducted. <Figure 4> illustrates the procedures to conduct T-test. The hypothesis of T-test is as follows :



〈Figure 4〉 T-Test Process

- Null Hypothesis : There is no difference in population mean between two groups.
- Alternative Hypothesis : There is difference in population mean between two groups.

The significance of hypothesis was determined by setting the level of significance of T-test as 0.02 Below <Table 5> illustrates the result of F-test and T-test for each genre. A symbol '○' indicates that null hypothesis is significant.

Looking into the result of F-test and T-test, the result of F-test is significant in all genres. Namely, hypothesis 'Variance of two groups is same.' was presented to be significant for F-test and hypothesis 'There is no difference in population mean of two groups.' was presented to be significant in all genres based on T-test.

〈Table 5〉 Result of F-Test and T-Test for Each Genre

genre	F-Test	T-Test
Action	○	○
Adventure	○	○
Animation	○	○
Biography	○	○
Comedy	○	○
Crime	○	○
Documentary	○	○
Drama	○	○
Family	○	○
Fantasy	○	○
Film-Noir	○	○
History	○	○
Horror	○	○
Music	○	○
Musical	○	○
Mystery	○	○
Romance	○	○
Sci-Fi	○	○
Sport	○	○
Thriller	○	○
War	○	○
Western	○	○

4.5 Recommendation for Movies

A group most similar with total ratings was found and whether or not the population

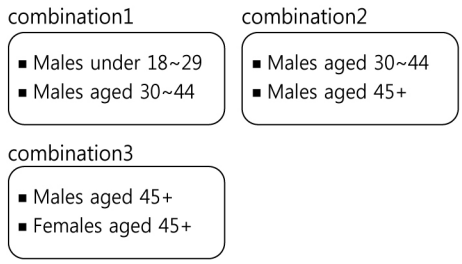
means of that group is same as total ratings group was examined. As a result, the population mean of each group extracted in each of 22 genres was same with that of total rating group. Therefore, a group most similar with total rating for each genre determined based on basic statistics on <Table 5> can be considered as a group that represents the group that illustrates total ratings.

Therefore, it is effective for a movie distributor to carry out targeted promotion for a representative group for each genre based on the result of above analysis rather than conducting marketing strategy that considers all age groups for each genre in preview or marketing stage. Moreover, it would be able to save resource including cost, manpower, time, etc.

5. T-Test for Combination of Age Group and Gender

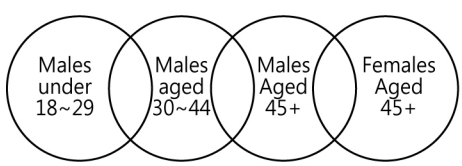
5.1 T-Test for Combination of Gender

The result of T-test for combination of gender presented to significant in 21 genres except Musical as you can see from the result of the first column in <Table 6>. It can be considered that similar ratings are given regardless of gender in all genres except Musical. 5.2 T-Test for Combination of Age Group and Gender. As you can see from the <Figure 5> which illustrates the result of T-test for combination of age group and gender, Males Aged 18~29/Females Aged 30~44 (C/F), Males Aged 30~44/Males Aged 45+ (E/G), and Males Aged 45+/Females Aged 45+ (G/H) combinations presented significant result in all genres. It can be considered that age group of two combinations give similar ratings regardless of genre. Therefore we can conclude that males under 18~29 group and males aged 30~44 groups-combination 1-have very similar user ratings and behaviors and other two combinations are the same.



〈Figure 5〉 Three Similar Rating Groups

Looking into the group category of 3 combinations, the redundancy is displayed. As you can see in <Figure 6>, 3 combinations are partially connected to one another. Therefore, category groups of Males under 18~29, Males aged 30~44, Males aged 45+, and Females aged 45+ that compose each combination can be classified as similar group.



〈Figure 6〉 Relations of Categories that Compose Combinations

〈Table 6〉 T-Test Result of the Combinations of Age Range and Gender

Combination Genre	M/W	A/B	A/C	A/D	A/E	A/F	A/G	A/H	B/C	B/D	B/E	B/F	B/G	B/H	C/D	C/E	C/F	C/G	C/H	D/E	D/F	D/G	D/H	E/F	E/G	E/H	F/G	F/H	G/H
Action	○	○	○	○	X	X	X	X	○	○	X	X	X	X	○	○	○	○	X	○	○	X	X	○	○	○	○	○	○
Adventure	○	○	○	○	X	○	X	X	X	○	X	X	X	X	○	○	○	○	○	X	○	X	X	○	○	○	○	○	○
Animation	○	○	○	○	X	○	X	X	X	○	X	X	X	X	X	○	○	○	○	X	○	X	X	○	○	○	○	○	○
Biography	○	○	○	○	X	X	X	X	○	○	○	○	○	○	○	X	○	○	○	X	X	X	X	○	○	○	○	○	○
Comedy	○	○	○	○	X	○	X	X	X	○	X	X	X	X	○	○	○	○	○	X	○	X	X	○	○	○	○	○	○
Crime	○	○	○	○	X	X	X	X	○	○	○	○	○	○	○	○	○	X	X	○	○	○	○	○	○	○	○	○	○
Documentary	○	○	○	○	○	○	○	○	X	○	X	○	X	○	○	○	○	○	○	○	○	○	○	○	○	○	X	○	○
Drama	○	X	○	○	○	○	X	○	X	X	X	X	X	X	○	○	○	X	○	X	○	X	X	X	○	○	X	X	○
Family	○	○	○	○	○	○	○	○	X	○	X	X	X	X	○	○	○	○	○	X	○	X	X	○	○	○	○	○	○
Fantasy	○	○	○	○	○	○	○	○	X	○	X	X	X	X	○	○	○	○	○	X	○	X	○	○	○	○	○	○	○
Film-Noir	○	○	○	○	○	○	○	○	○	○	○	○	○	X	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
History	○	X	○	○	X	X	○	○	○	X	○	○	○	○	○	○	○	○	○	X	X	○	○	○	○	○	○	○	○
Horror	○	○	X	X	X	X	X	X	○	○	○	○	○	○	○	○	○	○	○	○	○	X	X	○	○	○	○	○	○
Music	○	○	○	○	○	○	○	○	○	○	X	○	○	○	X	○	○	○	○	X	○	X	X	○	○	○	○	○	○
Musical	X	○	○	X	X	○	○	○	X	○	X	○	X	○	X	○	○	○	○	X	X	X	X	X	○	X	○	○	○
Mystery	○	○	○	○	X	X	X	X	○	○	○	○	○	○	○	○	○	○	○	X	○	X	○	○	○	○	○	○	○
Romance	○	○	○	○	X	○	○	○	○	○	X	○	○	○	○	○	○	○	○	X	○	X	X	○	○	○	○	○	○
Sci-Fi	○	○	○	○	X	X	X	X	○	○	X	X	X	X	○	○	○	○	X	○	○	X	X	○	○	○	○	○	○
Sport	○	○	○	○	X	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Thriller	○	○	○	○	X	X	X	X	○	○	X	○	X	X	○	○	○	○	○	X	○	X	○	○	○	○	○	○	○
War	○	X	○	○	X	X	X	X	X	X	○	○	X	X	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Western	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

- A : Males under 18
- B : Females under 18
- C : Males Aged 18~29
- D : Females Aged 18~29
- E : Males Aged 30~44
- F : Females Aged 30~44
- G : Males Aged 45+
- H : Females Aged 45+
- M : Man
- W : Woman
- ○ : Select a hypothesis
- X : Dismiss a hypothesis

Moreover, significant result was presented for Males under 18/Males Aged 18~29 (A/C), Males Aged 18~29/Males Aged 30~44 (C/E), Female Aged 30~44/Females Aged 45+ (F/H), Males Aged 30~44/Females Aged 45+ (E/H) and Males and Females (M/W) combinations in all genres except one genre respectively. Also result of Males Aged 18~29/Females Aged 30~44 (C/F), Males Aged 30~44/Males Aged 45+ (E/G) and Males Aged 45+/Females Aged 45+ (G/H) combinations is significant in all genres.

5.2 Result Based on Combination of Age Group and Gender for Each Genre

In combination of age group and gender, Western genre presented significant result in all combinations as you can see from <Table 6>. It can be considered that all age groups have given similar ratings in that genre on the average. On the other hand, insignificant result was presented for 16 combinations of age group in Drama genre thus it was presented to be a genre that was given most uneven ratings. Moreover, significant result was presented in all combinations of age group and gender except B/H combination in regards to Film-Noir genre and A/E combination in regards to sports genre.

6. Conclusion and Future Work

In this paper, we analyzed user movie ratings based on demographic information and found the representative group of each genre. This study searched the representative group most similar with total ratings through analysis on basic statistics by analyzing the rating data for each gender and age group and proposed to recommend a movie for that representative group after examining it with F-test and T-test. Males Aged 18~29 group was presented to be the group with most influence as it was presented to a group that represents the whole group in 13 genres. Then, the combination that presented significant result was found through combination of each age group and gender and it was classified as similar group. Moreover, 4 groups classified as similar group give similar ratings. Recommendation can be made by classifying it as a single category.

So, it is effective for a movie distributor to carry out targeted marketing for representative group for each gender and age group instead of preview and marketing for all age groups and gender. Also, 4 groups classified as similar group give similar ratings thus recommendation can be made by classifying it as a single category.

Therefore, this study is considerably significant as presenting utilization for film business as well as showing how to analyze demographic information on movie ratings on the

web. Limitation of this study lies in the fact that the size of a group that gives user ratings has not been considered. Each group has different sizes, so the result of skewness may be affected by majority group. In the future work, the size of group should be considered as a criterion of similarity test.

References

[1] Ahn, H., "A Hybrid Collaborative Filtering Recommender System Using a New Similarity Measure," Proceedings of the 6th WSEAS International Conference on Applied Computer Science, 2007.

[2] Bozidar, K., Dijana, O., and Nina, B., "Temporal Recommender Systems," Proceedings of the 10th WSEAS international conference on Applied computer and applied computational science, 2011.

[3] Choi, E., "Analysis of Future Growth in Korea Movie Industry," The Korea Contents Association, Vol. 8, No. 11, pp. 134-143, 2008.

[4] Dellarocas, C., Zhang, X., and Awad, N., "Exploring The Value of Online Product Reviews in Forecasting Sales : The Case of Motion Picture," Journal of Interactive Marketing, Vol. 21, No. 4, pp. 23-45, 2007.

[5] Evangelopoulos, N., "Text Mining for Customer Satisfaction Monitoring," Proceedings of 5th WSEAS Int. Conf. on Simulation, Modeling and Optimization, 2005.

[6] Ha, S., "Helping Online Customers Decide through Web Personalization," IEEE Intelligent Systems, Vol. 17, No. 6, pp. 34-43, 2002.

[7] John, S. and Pokorny, M., An Economic History of Film, Routledge, New York, 2005.

[8] Kazienko, P., "Product Recommendation in E-Commerce Using Direct and Indirect Confidence for Historical User Sessions," Proceedings of 7th International Conference on Discovery Science, 2004.

[9] Kazienko, P. and Kiewra, M., "Integration of Relational Databases and Web Site Content for Product and Page Recommendation," Proceedings of 8th International Database Engineering and Applications Symposium, IDEAS '04, 2004.

[10] Kazienko, P. and Kiewra, M., Intelligent Technologies for Inconsistent Knowledge Processing. Advanced Knowledge International, Adelaide, South Australia, 2004.

[11] Kazienko, P. and Kolodziejski, P., "Personalized Integration of Recommendation Methods for E-commerce," International Journal of Computer Science and Applications, Vol. 3, No. 3, pp. 12-26, 2006.

[12] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J., "Applying Collaborative Filtering to Usenet News," Communications Of The ACM, Vol. 40,

No. 3, pp. 77-87, 1997.

[13] Konstan, J. and Riedl, J., "Recommender Systems : From Algorithms to User Experience," User Model, User-Adapt, Interact, Vol. 22, No. 1-2, pp. 101-123, 2012.

[14] Krulwich, B., "Lifestyle Finder : Intelligent User Profiling Using Large-Scale Demographic Data," AI Magazine, Vol. 18, No. 2, p. 37, 1997.

[15] Lee, J., "How eWOM Reduces Uncertainties in Decision-making Process Using the Concept of Entropy in Information Theory," The Journal of Society for e-Business Studies, Vol. 16, No. 4, pp. 241-256, 2011.

[16] Lee, J. and Myoung, H., "Development of a Book Recommender System for Internet Bookstore using Case-based Reasoning," The Journal of Society for e-Business Studies, Vol. 13, No. 4, pp. 173-191, 2008.

[17] Mooney, R. and Roy, L., "Content-based book recommending using learning for text categorization," Fifth ACM Conference on Digital Libraries, 2000.

[18] Nagy, Z., "AJAX-Based Data Collection Method for Recommender Systems," Proceedings of the 16th WSEAS International Conference on Communications, 2012.

[19] Park, S. and Song, H., "Online reviews and Word-of-Mouth Effect on the movie : The Thrist," Korea Regional Communication Research Association, Vol. 10, No. 4, pp. 157-191, 2010.

[20] Pazzani, M., "A framework for collaborative, content-based and demographic filtering," Artificial Intelligence Review, Vol. 13, No.(5-6), 1999.

[21] Pilászy, I., Zibriczky, D., and Tikk, D., "Fast ALS-based Matrix Factorization for Explicit and Implicit Feedback Datasets," Proceedings of the fourth ACM conference on Recommender systems, 2010.

[22] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J., "GroupLens : An Open Architecture for Collaborative Filtering of Netnews," Proceedings of the 1994 ACM conference on Computer supported cooperative work, 1994.

[23] Rich, E., "User Modeling via Stereotypes," Cognitive Science, Vol. 3, pp. 329 354, 1979.

[24] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J., "Analysis of Recommendation Algorithms for E-Commerce," Proceedings of the 2nd ACM Conference on Electronic Commerce, 2000.

[25] Vassiliou, C., Stamoulis, D., and Martakos, D., "A Recommender System Framework combining Neural Networks and Collaborative Filtering," Proceedings of the 5th WSEAS Int. Conf. on Instrumentation, Measurement, Circuits and Systems, 2006.

[26] Vozalis, M. and Margaritis, K., "Using SVD and Demographic Data for the Enhancement of Generalized Collaborative Filtering," Information Sciences, Vol. 177,

No. 15, pp. 3017-3037, 2007.

[27] Wang, H. and Wu, C., "A Strategy-Oriented Operation Module for Recommender Systems in E-Commerce," Proceedings of the 9th WSEAS International Conference on Applied Informatics and Communications, 2009.

[28] Yang, W., Wang, Z., and You, M., "An improved collaborative filtering method for recommendations generation," 2004 IEEE International Conference on Systems, Man and Cybernetics, 2004.

저 자 소 개

배성문	(E-mail : bsm@gnu.ac.kr)
1993년	포항공과대학교 산업공학과 (학사)
1995년	포항공과대학교 산업공학과 (석사)
2001년	포항공과대학교 산업공학과 (박사)
2001년~2002년	LG CNS 연구개발센터
2003년~현재	경상대학교 산업시스템공학부 부교수
관심분야	소프트웨어 재사용, BPM, 시스템통합(SI) 등

이상천	(E-mail : sclee@gnu.ac.kr)
1989년	서울대학교 공과대학 산업공학과 (학사)
1991년	서울대학교 대학원 산업공학과 (석사)
1995년	서울대학교 대학원 산업공학과 (박사)
1991년~1998년	㈜대우전자부품
1998년~현재	경상대학교 산업시스템공학부 교수
관심분야	신뢰성공학, 확률모형, Multivariate Statistical Analysis

박종훈	(E-mail : icelatte@cu.ac.kr)
1997년	동국대학교 공과대학 산업공학과 (학사)
2000년	서울대학교 대학원 산업공학과 (석사)
2010년	서울대학교 대학원 산업공학과 (박사)
2002년~2004년	LG CNS 연구개발센터
2010년~2011년	서울대학교 산학협력단 박사후연구원
2011년~현재	대구가톨릭대학교 경영학과 조교수
관심분야	신뢰성공학, 확률모형, 품질경영, BPM 등