Journal of Data Science **11**(2013), 143-155

An Inference Model for Online Media Users

Narameth Nananukul
Asian University

Abstract: Watching videos online has become a popular activity for people around the world. To be able to manage revenue from online advertising an efficient Ad server that can match advertisement to targeted users is needed. In general the users' demographics are provided to an Ad server by an inference engine which infers users' demographics based on a profile reasoning technique. Rich media streaming through broadband networks has made significant impact on how online television users' profiles reasoning can be implemented. Compared to traditional broadcasting services such as satellite and cable, broadcasting through broadband networks enables bidirectional communication between users and content providers. In this paper, a user profile reasoning technique based on a logistic regression model is introduced. The inference model takes into account genre preferences and viewing time from users in different age/gender groups. Historical viewing data were used to train and build the model. Different input data processing and model building strategies are discussed. Also, experimental results are provided to show how effective the proposed technique is.

Key words: Data processing, demographics inference, inference model, logistic regression, profile reasoning.

1. Introduction

With the increase in internet broadcasting and web casting services, the ability to infer users' demographics such as gender and age group based on their historical usages is essential for effective targeted group advertising and content serving services. In the past the majority of broadcasting services were unidirectional where profiles of customers were not available, as a result providing customized contents (advertisement and/or programs) to customers was not possible. With the advance in online broadcasting environments where information such as usage history and customers' demographics are available, the ability to infer a customer profile based on existing information from customers with profiles is crucial for providing customized services to customers.

In general, an Ad server is a tool used by video publishers and advertisers to help with Ad management, campaign management and Ad trafficking. Typically, Ad servers maximize revenue for publishers and advertisers by first serving the highest paying Ad then follows by the second highest paying Ad that is available for each video viewer, and so on. The advertising campaign stops when the number of impressions required by the campaign is reached. Since different advertising campaigns target different demographics' groups of users, the ability to infer demographics information for users without profile is crucial to the success of Ad serving function. The demographics inference engine could help reduce waste of impressions in the advertising system. Figure 1 shows the relationship between an Ad server and an inference engine.

In Figure 1 the system collects users' profiles during registration process if users decide to register themselves to the system. The profiles can be created by using a provided user interface screen. The information is then stored in a profiles database. Users can view videos through a provided video player. The videos' contents are managed by a content server which retrieves media files from a videos' database. The displayed advertisement is managed by an Ad server. The advertisement is selected based on the demographic of the user (for users with profiles) or the inferred demographic of the user (for users without profiles) such that the revenue is optimized.

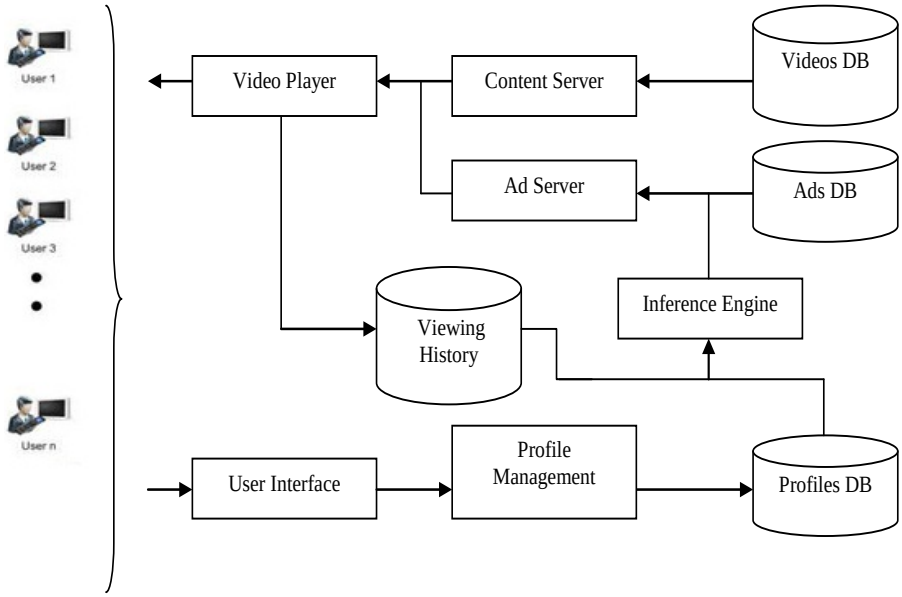

Figure 1: Ad serving system for online video provider

The objective of targeted online advertisement is to classify online media users into groups where each group contains customers with similar preferences.

Several collaborative filtering techniques have been used to infer users' favorite programs based on their historical usages. These techniques require profile information such as gender, job and age which can be collected through registration process. With users' historical viewing information customized advertisement and programs could be provided to customers based on their inferred favorite programs.

In this paper the focus is on building an inference model for inferring online media users' profiles by using logistic regression models. In general, the requirement of the inference is to generate demographic inference information for a list of targeted demographics. For each user, the output represents a set of inferred probabilities (IPs) that a user be a member of targeted demographics. The data used in this paper were sampled from the data from an online media provider that provides contents 24/7 with 61 genres.

At present, there are altogether 12 combinations of targeted demographics for online media users, 2 genders (male and female), and 6 age groups (<18, 18-24, 25-34, 35-44, 45-54, and 55+). Users are categorized into two different groups, known and inferred. A known user fits a particular demographic profile with probability 1, while an inferred user is not known with certainty to be a member of any demographic. For inferred users, when a sufficient amount of viewing data is available, their demographics inference can be determined by a user profile reasoning technique based on a logistic regression model.

The paper is organized as follows. Section 2 presents a literature review in the areas of collaborative filtering and profile reasoning techniques. Section 3 introduces the logistic regression model for the inference engine. Section 4 describes input data selection and data processing. Next, Sections 5 and 6 present model selection procedures and experimental results. Finally, Section 7 provides conclusions for the paper.

2. Literature Review

Although the scope of this research is to develop an inference model for predicting users' demographics, many related works are in the area of collaborative filtering. Billsus and Pazzani (1998) focused on the collaborative filtering problem where the users' rating matrix is sparse. In general, similarities between users that are calculated from a sparse rating matrix are inaccurate. As a result, accuracy of the predictions is poor. The authors proposed dimensionality reduction of a rating matrix with Singular Value Decomposition to capture the similarity in a reduced dimensional matrix.

Herlocker *et al.* (1999) provided an algorithmic framework that enhances the accuracy of predictions from the collaborative prediction process. The data were drawn from an analysis of historical data collected from an operational movie pre-

diction site. The authors developed a neighborhood-based prediction algorithm to perform automated collaborative filtering by using Spearman correlation as similarity weighting measure. Kurapati *et al.* (2001) developed a recommendation system for personal televisions. The personal televisions provide users with special devices called personal video recorders (PVRs). The proposed recommender system that help track users' preferences and aid users in choosing shows to record. The recommender system is a multi-agent TV recommender system that utilizes view history, preferences, and feedback information on specific shows to create adaptive agents and generate program recommendations for TV viewers. Burke (2002) proposed a hybrid recommender system that combines knowledge-based recommendation and collaborative filtering to recommend restaurants to users. The author shows that the semantic ratings obtained from the knowledge-based part of the system enhance the effectiveness of collaborative filtering.

Miyahara and Pazzani (2004) proposed an approach to collaborating filtering for web sites that recommend books, CDs and movies. Their approach is based on the simple Bayesian classifier which contains two variants of the collaborative filtering. One is user-based collaborative filtering, which makes predictions based on the users' similarities. The other is item-based collaborative filtering which makes prediction based on the items' similarities. They tested the model by using data from a database of movie recommendations. The empirical results show that their approaches outperform typical correlation-based collaborative filtering algorithms. Yu and Zhou (2004) developed an adaptive assistance to help personalizing interesting TV content to users. The proposed adaptive assistance observes users' viewing behaviors and updates user's profiles continuously, and then provides programs recommendations for different users according to their preferences. The novel aspect of the proposed system is that it evaluates the time it takes the system to learn new preferences after a preference is chosen. Sotelo *et al.* (2009) utilized semantic reasoning techniques to develop content recommendation for Digital TV and Personal Digital Recorders.

In the area of consumer clustering and targeted advertising using profile reasoning technique Bozios *et al.* (2001) presented consumer clustering and targeted advertising in a digital TV environment. Clusters of consumers are based on demographics, preferences, and analysis of the consumer interactions with the TV. Consumer data from the set top box (STB) are periodically transferred to the server where data mining techniques are applied to match consumer behavior with existing clusters. Then, the advertisements that match consumer's interests are displayed to consumer. Lim *et al.* (2008) proposed a user profile reasoning method for TV viewers. Their user profile reasoning is made in terms of genre preferences and TV viewing time for TV viewers in different genders and ages.

They proposed a multi-stage classifier as a profile reasoning algorithm. Historical viewing data from 2,522 users were used to build and test the system.

3. Logistic Regression

3.1 Background

The basic of logistic regression is based on the idea of predicting the odds of the outcome. In general the focus is on determining the odds of a user being in a targeted demographic group. Let the probability for a user being in a demographic group be p. Then the odds of an outcome is:

$$\text{odds} = (p/(1-p)). \tag{1}$$

The logistic regression can be represented as follows:

$$\begin{aligned} \text{Logit}(p) &= \log(p/(1-p)) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k, \end{aligned} \tag{2}$$

where $x_1, x_2, \cdots, x_k$ represent input variables to the model and $\beta_1, \beta_2, \cdots, \beta_k$, represent the regression coefficients for the input variables.

Although there is a general model of logistic regression called a multinomial logistic regression model which can be used to predict the probabilities of having different possible outcomes, predicting the inferred probability for each targeted demographic segment separately will lead to higher accuracy on prediction. The reason is because each targeted demographic segment will most likely be fitted with different set of regression coefficients in order to have the highest accuracy in prediction.

3.2 Advantages of Model Based on Logistic Regression

Logistic regression has several characteristics that are suitable for implementing the inference model as listed below:

- For regular linear regression, the IPs could become greater than one and less than zero. Those values are inadmissible for representing probability.

- Logistic function provides more realistic variance of the output, especially when there are two outcomes. In contrast to regular regression where the variance of the output is constant across values of inputs, the variance of the output from logistic function approaches zero as p approaches one or zero.

4. Input Data

The data used in the experiment were collected from an online media provider in the United States that provides contents 24/7. The data were collected during a period of one month which includes weekday and weekend.

4.1 Input Data Selection

Historical viewing data consist of several information that can be used as inputs for the regression model. For example, log data provide information on brand, genre, time of view and length of view for all contents viewed by users. A score system was used to determine what information should be selected as the regression model's inputs. In the experiment, 5 factors were considered, they consist of genre, brand, time of day, day of week and usage level. The scores are based on viewtime percentage categorized by gender, age and the factor in consideration. For each factor the scores (for gender and age) represent the total amount of absolute deviation of viewtime from the weighted average of viewtime (weighted by number of impressions) from all factor levels. In the experiment there are 61 levels for genre (number of genres), 62 levels for brand (number of brands), 3 levels for time of day (day, evening and night), 2 levels for day of week (weekday and weekend) and 3 levels for usage level (low, medium and high). Table 1 reports the scores for all factors considered.

Table 1: Scores from all input factors

	Factors	Gender	Age
1.	Genre	0.117	0.071
2.	Brand	0.114	0.058
3.	Time of day	0.025	0.028
4.	Day of week	0.002	0.013
5.	Usage level	0.028	0.075

From Table 1, genre and brand have high scores for both gender and age. Usage level has high score for age but not for gender. Time of day and day of week both have low scores for gender and age. Since the targeted segments are defined by gender and age, genre and usage level were selected as input factors. Note that brand was not selected because it has a one-to-one relationship with genre.

4.2 Data Processing

Building logistic regression model requires a training data set. The training

data consist of historical viewing data from known demographic users or users with registration information. The data consist of output training data and input training data. The output training data consist of a vector of binary variables (targeted demographic vector). The variable is set to 1 if the associate user belongs to the targeted demographic, set to 0, otherwise. The input training data consist of an array of genre vectors and a usage level vector.

A genre vector contains view percentages for all users. For a user genre view percentage can be determined by dividing user's genre view time by user's total view time. The usage level vector contains usage time from all users.

To be able to weigh different types of inputs equally the input data were standardized by encoding input vectors into 0/1. Each vector is converted to a standardized vector by setting the data to 0 if data is below the "threshold" value, and converted to 1, otherwise. The threshold for each vector was selected such that the standardized vector has the highest correlation with the targeted demographic vector.

5. Model Selection

5.1 Model's Input Variables

Using too many input variables will generally cause overfitted issue, on the other hand using too few input variables will deteriorate model accuracy. The input variables were chosen based on their statistical significances (p-values). Those variables with p-values at least 0.1 (90% significant level) were selected as model's input variables.

The logistic regression fitting process was implemented by using logistic regression module provided by R statistical software package. By default the statistical significance of each coefficient in the model is computed by using a Wald test (Everitt and Hothorn 2006, Chapter 6).

5.2 Choosing Interaction Terms

One of the assumptions of the logistic regression model is that each input has an independent effect on the output variable. Input variables do not have interaction or joint effect unless it is specified by including interaction terms into the model. In general, an interaction term represents the effect of one input variable on the output variable as a function of another input variable. For example, users that love to watch both sports and action movies have more chance to be male.

The pair-wise interaction terms between input variables were chosen based on AIC (Akaike information criterion) values. The model with the lowest AIC

value was selected. AIC values were generated by using a logistic regression module provided by R. R provides a stepwise regression that takes a series of steps by either deleting a term already in the model or adding a term from a list of candidates for inclusion. Selection of terms for deletion or inclusion is based on AIC values. R defines the AIC value as 2*(maximized log-likelihood) + 2*(number of parameters) (Fox 2002, Chapters 4 and 6). The procedure stops when the AIC value cannot be improved.

6. Experimental Results

In this section experimental results based on the proposed approach are provided. The targeted segment (male, age 45-54) was chosen for the experiment because it contains the most number of impressions. In this targeted group there are altogether 9862 users with registration information. 80% of the data were used as training data while 20% of the data were used as test data. The model accuracy is measured by the accuracy index. The detail on accuracy index calculation is presented next.

6.1 Accuracy Index

The accuracy of the logistic regression model is measured by comparing the average of users' IPs and the actual mean of users' demographic. To be able to take into account the accuracy from different ranges of IPs the IPs were divided into intervals where the break points are at $0, 0.1, \cdots, 0.9, 1.0$ and the mean of targeted demographic. In general, we are interested in IPs that are higher than the mean of targeted demographic, as a result all brackets range below the mean of targeted demographic were removed from the calculation. Also, brackets that have less than 10 records were not defined separately but were merged and redefined as a new bracket. The average inferred probability and prediction error for each interval were then calculated. The accuracy index is the ratio of the weighted average of IPs and absolute prediction errors from every interval. The example in Table 2 illustrates how the accuracy index can be calculated. Note that in this example the mean of target demographic is 0.16.

Table 2: Example of results from the inference model

Upper Range	Lower Range	Avg	Age IPs	Deviation	#Records
	≥ 0.4	0.375	0.468	0.249	8
< 0.4	≥ 0.3	0.414	0.337	0.188	41
< 0.3	≥ 0.2	0.216	0.234	0.084	250
< 0.2	≥ 0.16	0.21	0.177	0.157	328

There are 6 columns in the table where the first two columns (Upper Range and Lower Range) represent the upper and lower limit of the inferred probabilities. The average of users' demographics (Avg) and users' IPs (Avg IPs) are shown in column 3 and 4, respectively.

Column 5 is the percent deviation (absolute deviation) of Avg IPs from Avg. The last column contains number of records in each interval. From the data the product sum of Avg IPs and #Records is 134.27 and the product sum of Deviation and #Records is 82.19. As a result the accuracy index is 134.27/82.19 = 1.63.

6.2 Comparing Results from Different Input Formats

Different formats of input data were used to determine the best way to represent input data for the inference model. The first input data format consists of 0/1 encoded data for genres. The genre vector can be prepared by setting the data to 1 if user's view time for the genre is greater than 0,set to 0, otherwise. This is the most basic format for representing input data. Although it is easy to implement, the result accuracy is not very good because users with similar viewing habit but different usage levels are treated the same way. The result is shown in Table 3.

Table 3: Result when 0/1 encoded-genre information is used as input data (case 1)

Upper Range	Lower Range	Avg	Age IPs	Deviation	#Records
	≥ 0.4	0.143	0.444	2.106	7
< 0.4	≥ 0.3	0.37	0.339	0.084	54
< 0.3	≥ 0.2	0.23	0.234	0.018	300
< 0.2	≥ 0.16	0.133	0.178	0.333	345
				Acc Index	1.1

In order to capture users' viewing habit from the data the second input data format was introduced. In this format genre-ratio is used to create genre vectors. Genre-ratio for each user can be generated by first calculating total view time for each genre, then divide it by total view time. Then, genre vector can be generated by setting the data to 1 if the genre-ratio for the genre is greater than a threshold, set to 0, otherwise. The thresholds are used to distinguish between high usage users and low usage users for every genre, they were chosen such that each genre vector has the highest correlation with the targeted demographic vector.

Using genre-ratio help normalize users' genre-viewtime by their total view time. The result is summarized in Table 4. The result shows that this method

can improve about 18 percent of the accuracy index. Also, notice that on average Avg IPs are higher but the result does not show significant improvement in deviation from all brackets. This reveals the drawback of using genre-ratio data format where users with very low usage (with random behavior) can degrade the accuracy.

Table 4: Result when 0/1 encoded genre-ratio with threshold is used as input data (case 2)

Upper Range	Lower Range	Avg	Age IPs	Deviation	#Records
	≥ 0.4	0.4	0.507	0.268	5
< 0.4	≥ 0.3	0.414	0.342	0.175	41
< 0.3	≥ 0.2	0.21	0.236	0.121	252
< 0.2	≥ 0.16	0.226	0.178	0.211	274
				Acc Index	1.29

To be able to take into account users' usage levels a new dimension for input data was introduced. This new dimension is called "usage level" vector. Usage level vector can be generated by setting the data to 1 if user view time is greater than a threshold, set to 0, otherwise. The threshold that results in a usage level vector that has the highest correlation with the target demographic vector was chosen. The result in Table 5 shows that adding usage level information improves accuracy index by 26 percent.

Table 5: Result when 0/1 encoded genre-ratio and usage level vector with threshold are used as input data (case 3)

Upper Range	Lower Range	Avg	Age IPs	Deviation	#Records
	≥ 0.4	0.375	0.468	0.249	8
< 0.4	≥ 0.3	0.414	0.337	0.188	41
< 0.3	≥ 0.2	0.216	0.234	0.084	250
< 0.2	≥ 0.16	0.21	0.177	0.157	328
				Acc Index	1.63

To be able to screen out users with very low usage, users with view time less than "minimum usage threshold" were removed from training data set. Again the a minimum usage threshold was selected such that the truncated usage vector has the highest correlation with the truncated target demographic vector. The result after screening out low usage users is summarized in Table 6. While the deviations for high inferred probability brackets (≥ 0.4) and [0.3,0.4) increase, the result shows a significant decrease in deviations for brackets [0.2,0.3) and

[0.16,0,2). Since the number of users that falls into brackets (≥ 0.4) and [0.3,0.4) is small, the result accuracy index is improved by 180 percent.

Table 6: Result when 0/1 encoded genre-ratio and usage level with threshold are used as input data after removing users with low usage (case 4)

Upper Range	Lower Range	Avg	Age IPs	Deviation	#Records
	≥ 0.4	0.25	0.442	0.769	8
< 0.4	≥ 0.3	0.25	0.337	0.347	44
< 0.3	≥ 0.2	0.234	0.235	0.005	248
< 0.2	≥ 0.16	0.182	0.178	0.019	297
				Acc Index	4.58

Figures 2 and 3 show comparisons of average inferred probabilities and deviations from case 1 to case 4. From Figure 2 notice that the Avg IPs are at the same levels for most of the ranges of IPs except for high IPs bracket (≥ 0.4). The results show that the input data formats do not have significant effect on the Avg IPs.

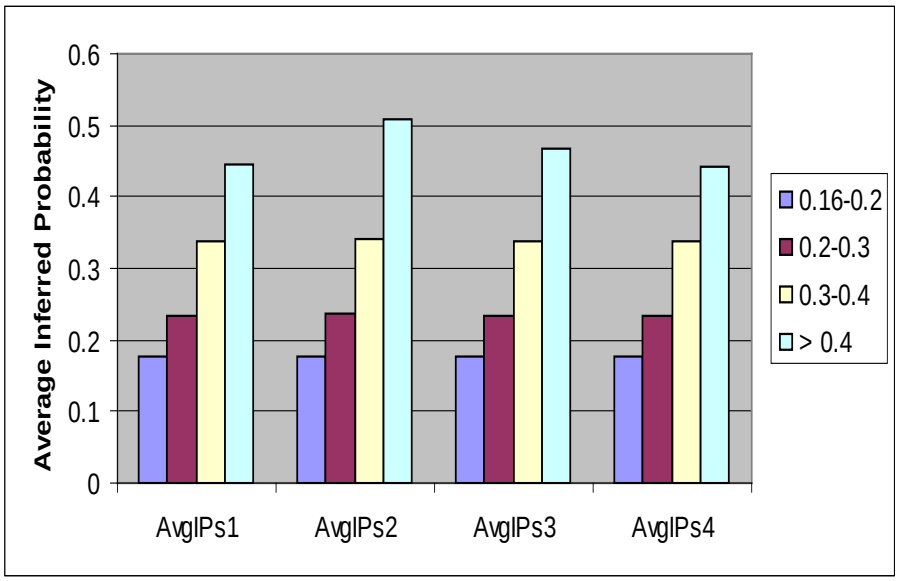

Figure 2: Comparison of average inferred probabilities from case 1 to case 4

As shown in Figure 3 on average the deviations from low to medium ranges of IPs ([0.16-0.2) and [0.2,0.3)) keep decreasing from case 2 to case 4. Since the majority of users fall into these two brackets the result shows that the input data format does help decrease the deviations from case 2 to case 4.

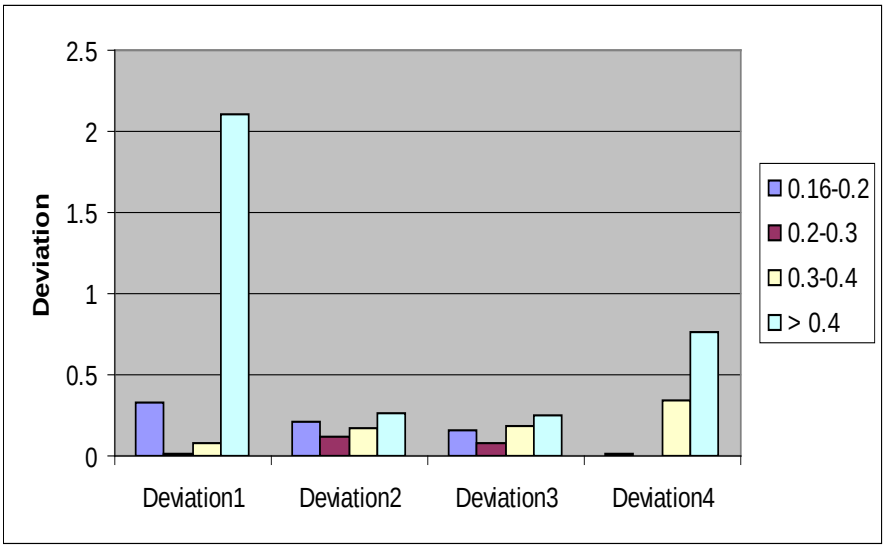

Figure 3: Comparison of deviations from case 1 to case 4

7. Summary

In this paper an inference model based on a profile reasoning technique was proposed. The inference model is based on a logistic regression model that uses historical viewing data from users as input data. Different input data formats were considered. From the experiment it shows that using genre-ratio in combination with usage level information can help improve the inference accuracy. Using thresholds to screen out negligible data when creating usage level vectors can reduce randomness in the data.

References

Billsus, D. and Pazzani M. J. (1998). Learning collaborative filters. In *Proceedings of the 15th International Conference on Machine Learning*, 46-54. San Francisco, California.

Bozios, T., Lekakos, G., Skoularidou, V. and Chorianopoulos K. (2001). Advanced techniques for personalized advertising in a digital TV environment: the iMEDIA system. In *Proceedings of The E-business and E-work Conference*, 1025-1031. Venice, Italy.

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

Everitt, B. and Hothorn, T. (2006). *A Handbook of Statistical Analyses Using R.* Chapman and Hall/CRC, Boca Raton, Florida.

Fox, J. (2002). *An R and S Plus Companion to Applied Regression.* Sage Publications, Thousand Oaks, California.

Herlocker, J., Konstan, J., Borchers, A. and Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. In *Proceedings of 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 230-237. Berkeley, California.

Gutta, S., Kurapati, K., Lee, K. P., Martino, J., Milanski, J., Schaffer, J. D. and Zimmerman, J. (2000). TV Content Recommender System. *Proceedings of the 17th National Conference on Artificial Intelligence, Austin*, 1121-1122. AAAI Press/The MIT Press, Texas.

Lim, J., Kim, M., Lee, B., Kim, M., Lee, H. and Lee, H. K. (2008). A target advertisement system based on TV viewer's profile reasoning. *Multimedia Tools and Applications* **36**, 11-35.

Miyahara, K. and Pazzani, M. J. (2004). Collaborative filtering with the simple Bayesian classifier. In *Proceeding of the 6th Pacific Rim International Conference on Artificial Intelligence* (*PRICAI 2000*), 679-689. Springer-Verlag, Berlin, Heidelberg.

Sotelo, R., Blanco-Fernandez, Y., Lopez-Nores, M., Gil-Solla, A. and Pazos-arias, J. (2009). TV program recommendation for groups based on muldimensional TV-anytime classifications. *IEEE Transactions on Consumer Electronics* **55**, 248-256.

Yu, Z. and Zhou, X. (2004). TV3P: An adaptive assistant for personalized TV. *IEEE Transactions on Consumer Electronics* **50**, 393-399.

Received June 25, 2012; accepted September 28, 2012.

Narameth Nananukul
Department of Technology Management
Faculty of Engineering and Technology
Asian University
Chon Buri, 20150, Thailand
naramethn@asianust.ac.th