Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 15(1): 43-48, Mar. 2017 **Regular paper**

Bayesian Approach to Users' Perspective on Movie Genres

Artem A. Lenskiy* **and Eric Makita,** ***Member, KIICE***

Department of Information and Communication Engineering, Korea University of Technology and Education, Cheonan 31253, Korea

Abstract

Movie ratings are crucial for recommendation engines that track the behavior of all users and utilize the information to suggest items the users might like. It is intuitively appealing that information about the viewing preferences in terms of movie genres is sufficient for predicting a genre of an unlabeled movie. In order to predict movie genres, we treat ratings as a feature vector, apply a Bernoulli event model to estimate the likelihood of a movie being assigned a certain genre, and evaluate the posterior probability of the genre of a given movie by using the Bayes rule. The goal of the proposed technique is to efficiently use movie ratings for the task of predicting movie genres. In our approach, we attempted to answer the question: "Given the set of users who watched a movie, is it possible to predict the genre of a movie on the basis of its ratings?" The simulation results with MovieLens 1M data demonstrated the efficiency and accuracy of the proposed technique, achieving an 83.8% prediction rate for exact prediction and 84.8% when including correlated genres.

Index Terms: Item category prediction, Multivariate Bernoulli event model, Naive Bayes classification

I. INTRODUCTION

Nowadays, Internet users are no longer simply considered consumers of information; they are also believed to be active sources that generate a large volume of data online. Consequently, the amount and the diversity of information on the Internet increase exponentially. The whole body of information that confronts the users online can cost them more time and effort without any guarantee of finding what they are looking for. Aiming to solve these problems, researchers in academia and/or industries have suggested the use of recommender systems [1] that overcome the information overload by facilitating search and access to information, thereby providing users with relevant items in the shortest possible time. In this context, items can be of any kind, namely, a movie to watch, a soundtrack to listen to, or a webpage to visit. Among the widely-proposed recommendation techniques, content-based filtering [2, 3] and collaborative filtering [4, 5] have been the most widely used in the literature [6]. Content-based filtering is done under the assumption that users' future preferences are similar to their past preferences, while collaborative filtering is done under the assumption that if two users had similar preferences in the past, they will have similar preferences in the future. Between the two, collaborative filtering is more widely used and therefore, attracts more interest from researchers [7, 8]. Collaborative filtering techniques are rating-oriented and involve the participation of a large number of users who provide fewer ratings than the items they consume. Taking this into consideration, questions such as how to alleviate the data sparsity while increasing the recommendation accuracy are the main concerns in the related work.

Recently, some approaches considering factors outside

Received 09 January 2017, **Revised** 09 January 2017, **Accepted** 25 January 2017
***Corresponding Author** Artem A. Lenskiy (E-mail: lensky@koreatech.ac.kr, Tel: +82-41-560-1165)
Department of Information and Communication Engineering, Korea University of Technology and Education, 1600, Chungjeol-ro, Byeongcheon-myeon, Dongnam-gu, Cheonan 31253, Korea.

Open Access **http://doi.org/10.6109/jicce.2017.15.1.43** print ISSN: 2234-8255 online ISSN: 2234-8883

users' ratings have been proposed in the literature [9]. Since recommender systems can naturally be applied in various fields where items are categorized, their associated datasets provide not only user ratings but also item categories. On the basis of this information, many recommender system extensions have been created.

In this paper, we propose a movie genre prediction model based on user ratings. We apply a multivariate Bernoulli model to estimate likelihoods that are used in the naïve Bayes rule to predict movie genres. We also calculate the genre correlations to check whether an incorrectly predicted genre is correlated with the correctly predicted one. In general, a recommender predicting an item's category is important in that it can complement the item's categories assigned by a human expert (i.e., a movie director), thereby increasing user satisfaction by providing surprising recommendations.

The proposed approach involves predicting movie' genre information on the basis of users' feedback, namely users' perception of the content genre information.

The remainder of this paper is organized as follows: Section II presents the literature review related to item-genre information. In Section III, we describe the data model and the mechanism of the proposed approach. Section IV contains the performance studies. In Section V, we elaborate on the results and discuss our future work. Finally, in Section VI, we summarize our work.

II. RELATED WORK

A considerable amount of research related to item content genres or categories has been conducted in the past.

One study proposed a movie recommender system that enhances the accuracy and overcomes the traditional recommendation by factorizing the user-genre matrix instead of the user-item matrix [10]. The factorized user-genre matrix model was used for discovering latent factors from genres in order to enrich user profiles. In [11], content-based filtering utilizing user category-based filtering was proposed to overcome one of the major issues of recommender systems termed as item cold start. Item cold start refers to new items that have not received sufficient feedback from users and thus, could decrease the accuracy of the recommendation. Another example of category-based recommendation is proposed in [12], where the authors presented a framework called the semantic enhanced personalizer for overcoming recommender system problems such as cold start and sparsity. The authors in [13] proposed a recommender system approach that uses genre information to address not only the coverage of the recommender system algorithms but also the redundancy. Most of the related works focus on designing a new approach to identify similarities between users, while the prediction of movie genres remains understudied to the best of our knowledge. However, it could play an important role in recommending novel items to a user.

III. BERNOULLI EVENT MODEL

The proposed method applies the well-known multivariate Bernoulli model for calculating the conditional probability of a movie being of a particular genre. To describe our idea clearly, we initially give some definitions used in this paper: User set: *U*, movie set: *M*, genre set: *G*, and rating set: *R*.

A movie $m \in M$ is characterized by a binary feature vector, whose components are set to 1 if the corresponding user u rated the movie m; otherwise, these components are set to zero.

$$v_{u,m} = \begin{cases} 1, & \text{if } u \text{ rated movie } m \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

With such a movie representation, we make the assumption that *user ratings are independent.* Then, the probability of a movie m being given its genre g is computed as the product of the user probabilities of the movie attribute values over all the user perceptions of genre information as follows:

$$P(m|g) = \prod_{u \in U} P(u|g)^{v_{u,m}} \left(1 - P(u|g)\right)^{1-v_{u,m}}. \quad (2)$$

A movie can be seen as a collection of multiple independent Bernoulli experiments, one for each user in the user set U with the probabilities for each of these user events defined by each component $P(u|g)$, i.e., the user's preferences towards genre g.

Rewriting (2) slightly differently, we obtain the following:

$$P(m|g) = \prod_{u \in U} P(u|g)^{v_{u,m}} \prod_{u \in U} Q(u|g)^{1-v_{u,m}}, \quad (3)$$

where $Q(u|g) = 1 - P(u|g)$ denotes the probability of not rating a movie of genre g. The first product represents the product of the user preferences and covers all users who rated a movie. Higher values of $P(u|g)$ result in a greater value of the product. The second product represents the product of the users' dislike towards a genre and covers all users who did not rate the movie. Greater values of $Q(u|g)$ result in a greater value of the product. In other words, $P(m|g)$ is greater if users who like genre g rated the movie m or they did not rate the movie m and they disliked genre g.

With a large number of users, the product (2) decreases quickly to zero. Thus, to address this underflow problem, we use the following log transformation:

$$\log\big(P(m|g)\big) = \sum_{u\in U} v_{u,m} \log\left(\frac{P(u|g)}{1-P(u|g)}\right) + \sum_{u\in U} \log\big(1-P(u|g)\big). \quad (4)$$

The last sum on the right-hand side can be precomputed, and the first term ranges only over the users that rate the current movie m, i.e., $v_{u,m} = 1$.

The probability $P(u|g)$ defines the user preferences towards different genres and can be estimated as follows:

$$P(u|g) = \frac{1 + \sum_{m\in M} v_{u,m} P(g|m)}{|U| + \sum_{m\in M} P(g|m)}. \quad (5)$$

Eq. (5) describes the probability that given the genre g, user u rates a movie of this genre. The probability $P(g|m)$ is 1 if m is marked as genre g; otherwise, it is 0. If movie m simultaneously belongs to N genres, the probability of $1/N$ is assigned. Eq. (5) involves taking a sum over $v_{u,m} \cdot P(g|m)$, if a user u did not rate any movies of a genre g then the whole sum $\sum_{m\in M} v_{u,m} P(g|m)$ is zero. However, just because a user does not watch a movie of a particular genre g in the training dataset does not mean that she cannot watch any movie of that genre. Even if user u did not rate a movie of genre $G = g$ in the training set, we would still like to have $P(u|g) > 0$ at it can appear during testing. Taking into account that probabilities must sum to 1, if users who did not rate a movie have probabilities near zero, then those users who watch must have higher probabilities. One efficient way to alleviate this problem, called the zero-probability problem, is to remove a small amount of probability allocated to those users who watched the movie in the genre and distribute this across those users who did not watch the movie. To this end, we use Laplace's law of succession [14] and add a count of one to each user type in the numerator. The denominator is increased to take into account the $|U|$ extra users arising from the "add 1" term, assuring that the probabilities are still normalized.

Given Eq. (5), $P(m|g)$ is evaluated on the basis of the movies in the training dataset, and then, the prediction of a movie m into genre g is computed on the basis of the posterior probability of each genre given the evidence of the test movie, and selecting the genre g with the highest probability according to Bayes' rule as follows:

$$P(g|m) = \frac{P(m|g) \cdot P(g)}{P(m)}, \quad (6)$$

where the prior probability is computed by simply counting the number of a movie within genre g over the total number of movies:

$$P(g) = \frac{\sum_{m\in M} P(g|m)}{|M|}. \quad (7)$$

To avoid the underflow problem, the prediction of a movie genre is performed in the log scale, as follows:

$$\hat{g} = \arg\max_{g} \left[\sum_{u\in U} v_{u,m} \log\left(\frac{P(u|g)}{1-P(u|g)}\right) + \sum_{u\in U} \log\big(1-P(u|g)\big) + \log\left(\sum_{m\in M} P(g|m)\right)\right]. \quad (8)$$

Here, we omit $-\log(P(m))$ and $-\log|M|$ as they do not depend on g.

IV. EXPERIMENTAL RESULTS

A. Dataset

We performed our experiments on the MovieLens 1M dataset [15], which contains 3,942 movies, 6,040 users, and approximately 1 million ratings that range from 1 to 5. After preprocessing and removing movies that did not have any ratings, we kept 3,706 movies, having a total of 1,000,209 ratings. Then, 18 movie genres were selected and each movie was assigned to at least one genre by movie experts. We carried out the experiment by dividing the dataset into two, a training set and a test set. During the training phase, rows of the genre matrix and columns in the rating matrix that correspond to the testing set were removed. Thus, the users' preference models were built using only a portion of the available rated items with the known genres selected for testing.

B. Evaluation

In this section, we demonstrate the correctness of the hypothesis that user preferences can be used for predicting movie genres.

Movies were selected randomly in both of our training and testing approaches. To test the prediction accuracy of the proposed approach, we used {1%, 5%, 10%, …, 75%, 80%} portions of the whole set as the training set. As for the testing set, we always kept it fixed at 20% of the whole dataset. None of the items of the testing dataset were included in the training. For every training size, we repeated the process of randomly selecting training samples 100

times. The prediction was considered to be successful if the predicted genre belonged to the set of true genres. Fig. 1 shows portions of movies with 1, 2, 3, 4, and 5 genres. Half of all the movies were assigned to only one genre.

When the prediction was incorrect, we checked whether the incorrectly predicted genre correlated with the true movie's genre. If it was correlated $(cor > 0.1)$, we accepted this prediction as a prediction of a similar genre. Fig. 2 shows the covariance matrix estimated for the genres. The red dots represent the correlated genres.

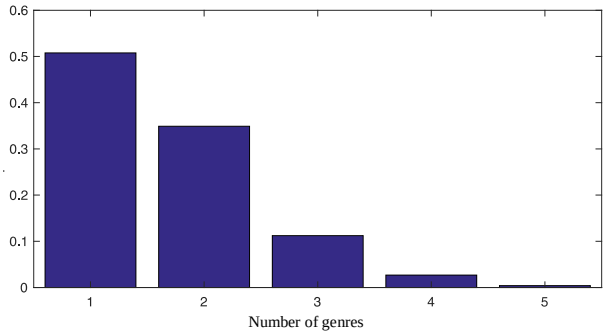

Fig. 1. Estimated probability of the number of genres per movie i.e. 50% of movies are assigned only one genre.

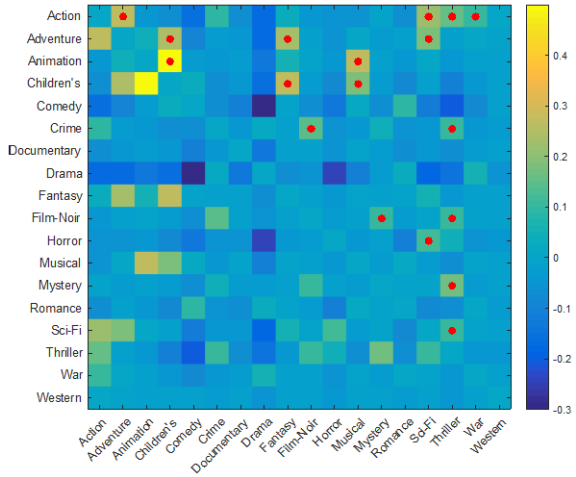

Fig. 2. Genre covariance matrix.

Table 1. Prediction rate

	Exact prediction	Including correlated genres
Prior	24.12 ± 1.3	24.12 ± 1.3
$r = 1$	71.70 ± 1.5	73.04 ± 1.4
$r = 2$	73.40 ± 1.4	75.20 ± 1.4
$r = 3$	79.50 ± 1.3	81.30 ± 1.3
$r = 4$	79.80 ± 1.4	81.20 ± 1.4
$r = 5$	76.00 ± 1.4	78.20 ± 1.4
Combined	83.80 ± 1.3	84.80 ± 1.4

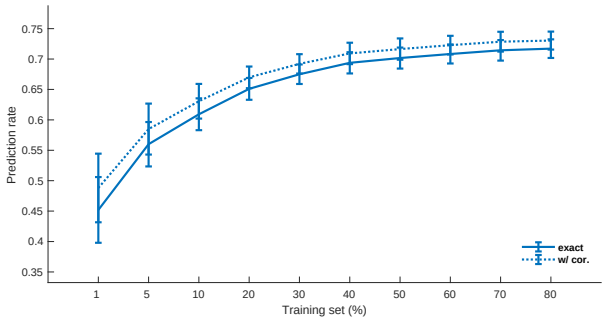

Fig. 3. Prediction accuracy based on rating $r = 1$.

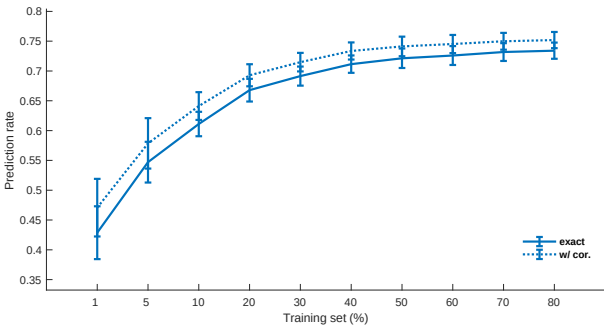

Fig. 4. Prediction accuracy based on rating $r = 2$.

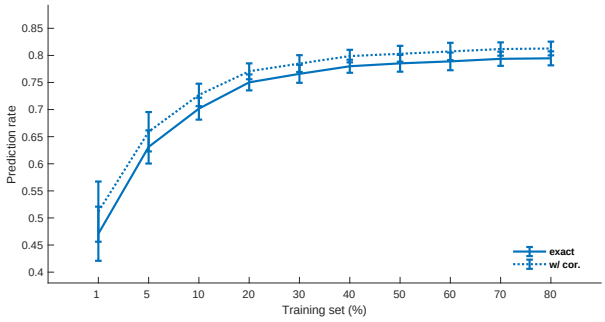

Fig. 5. Prediction accuracy based on rating $r = 3$.

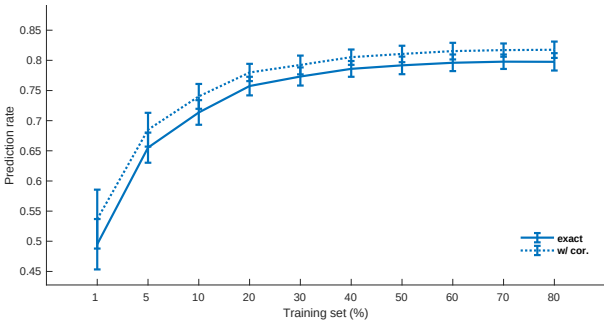

Fig. 6. Prediction accuracy based on rating $r = 4$.

The plots in Figs. 3–7 indicate that our movie category prediction based on the Bayesian model presented in Section III was effective in predicting genres based on all the ratings. With an increase in the size of the training

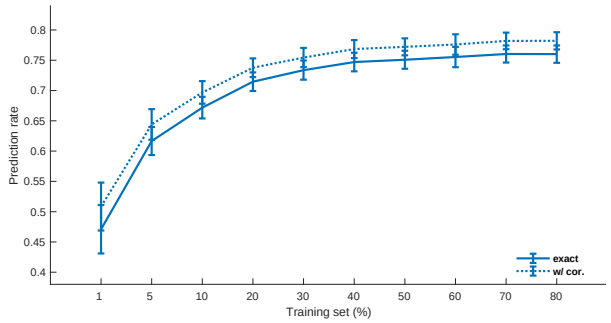

Fig. 7. Prediction accuracy based on rating $r = 5$.

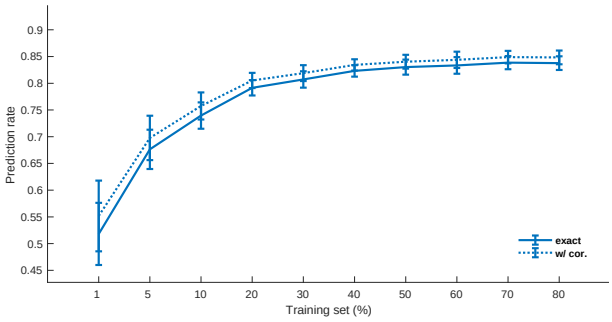

Fig. 8. Prediction accuracy based on combined ratings.

dataset, the prediction accuracy increased. Table 1 presents the prediction accuracy when 80% of the whole dataset was used for the training. The results show that both rating r = 3 and r = 4 produced higher accuracy values as they were the most frequently given ratings. The prediction accuracy rate for genres with the correlated genres included was always higher than the exact prediction rating for all of the ratings because the algorithm referred to the covariance matrix (Fig. 2) to check whether the incorrectly predicted genre was correlated to one of the true genres, and thus, the set of correct genres was enlarged.

It is interesting to note that for rating r = 3, only 10% of the data used for training was sufficient to achieve a 70% prediction accuracy.

The next was to combine all the ratings, and see whether a better accuracy could be achieved. To predict a genre on the basis of all the ratings, we simply summed the posterior probabilities for all ratings and chose a genre that maximized the sum:

$$\hat{g} = \arg\max_{g} \sum_{r \in R} \log(P_r(g|m)), \qquad (9)$$

where the terms in the sum were evaluated combining (4) and (6).

Combining the ratings improved the prediction, achieving an 83.8% prediction rate. The results are presented in Fig. 8.

V. DISCUSSION

We also compared the prediction accuracy using user preferences with a simple prediction based on prior probability. The highest prior probability corresponded to the genre drama, and hence, all movies were predicted to be dramas. The prediction accuracy using only a prior was only 24.1%. Because the most popular genre (drama) was not correlated with any other genres (Fig. 2), the prediction including the correlated genres, was equal to the accuracy of the exact prediction. As observed from Fig. 8, even taking into account the preferences of only 1% of all users, i.e., 60 users, resulted in a prediction accuracy of more than 50%.

Another interesting observation was the difference between the exact prediction accuracy and the prediction including the correlated genres. In the case of combined ratings (Fig. 8), the difference was smaller than in cases of predictions based on each rating separately. This may be attributed to the fact that combined ratings contain more information about the movies, and thus, the correlation does not provide additional information. This effect will be investigated in our future work.

It is clear that user preferences represent a rich source of information about movie genres, and in 83.8% of cases, users see a movie's genre just as a movie director labeled it. However, in the remaining 16.2% of cases, user perceptions did not match the genres assigned by directors. Although this could be considered an error in movie genre prediction, it can also be interpreted as an advantage of recommender systems, as these genres are seen as complementing genres and thus, can provide a surprising recommendation. This statement, although plausible, requires further study and testing.

VI. CONCLUSIONS

Over the last decade, movie genre prediction has been successfully applied in various domains such as social networking, online movie viewing websites, and e-commerce. Thus far, most of the techniques that have been proposed and reported in the literature do not take into account the perception of the users about the content genre information. In this work, we proposed an approach that expands the traditional movie classification algorithms by predicting the genre of a movie under evaluation instead of using user ratings of the watched movies. This approach can be easily generalized from movies to other items and their corresponding categories.

To show the correctness of our approach, we conducted an experimental study using the MovieLens 1M dataset. The experimental results showed that predicting the genre of a movie under evaluation can achieve an accuracy of more

than 50% on the basis of only 1% of the training set of the users' combined ratings and of 83.8% when 80% of the whole set was taken as the training set. This finding is deemed valuable in many applications in practice. For instance, it can complement the genres given by experts.

REFERENCES

[1] L. Lu, M. Medo, C. H. Yeung, Y. C. Zhang, Z. K. Zhang, and T. Zhou, "Recommender systems," *Physics Reports*, vol. 519, no. 1, pp. 1-49, 2012.

[2] T. Bogers and A. Van den Bosch, "Collaborative and content-based filtering for item recommendation on social bookmarking websites," in *Proceedings of the 2009 ACM Conference on Recommender Systems (RecSys)*, New York, NY, pp. 9-16, 2009.

[3] J. Basilico and T. Hofmann, "Unifying collaborative and content-based filtering," in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

[4] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in Artificial Intelligence*, vol. 2009, article no. 421425, pp. 1-19, 2009.

[5] F. Braida, C. E. Mello, M. B. Pasinato, and G. Zimbrao, "Transforming collaborative filtering into supervised learning," *Expert Systems with Applications*, vol. 42, no. 10, pp. 4733-4742, 2015.

[6] D. H. Park, H. K. Kim, I. Y. Choi, and J. K. Kim, "A literature review and classification of recommender systems research," *Expert Systems with Applications*, vol. 39, no. 11, pp. 10059-10072, 2012.

[7] Q. Liu, E. Chen, H. Xiong, C. H. Ding, and J. Chen, "Enhancing collaborative filtering by user interest expansion via personalized ranking," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 1, pp. 218-233, 2012.

[8] S. Vargas, L. Baltrunas, A. Karatzoglou, and P. Castells, "Coverage, redundancy and size-awareness in genre diversity for recommender systems," in *Proceedings of the 8th ACM Conference on Recommender Systems*, Foster City, CA, pp. 209-216, 2014.

[9] Z. Huang, H. Chen, and D. Zeng, "Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering," *ACM Transactions on Information Systems*, vol. 22, no. 1, pp. 116-142, 2004.

[10] M. G. Manzato, "Discovering latent factors from movies genres for enhanced recommendation," in *Proceedings of the 6th ACM Conference on Recommender Systems*, Dublin, Ireland, pp. 249-252, 2012.

[11] M. Sollenborn and P. Funk, "Category-based filtering and user stereotype cases to reduce the latency problem in recommender systems," in *Advances in Case-Based Reasoning*, Heidelberg: Springer, pp. 395-405, 2002.

[12] R. Tilwani and S. Tiwari, "Implementation of category based recommendation module of SEP architecture using PBTA," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 10, pp. 635-642, 2013.

[13] S. M. Choi, S. K. Ko, and Y. S. Han, "A movie recommendation algorithm based on genre correlations," *Expert Systems with Applications*, vol. 39, no. 9, pp. 8079-8085, 2012.

[14] Y. H. Li and A. K. Jain, "Classification of text documents," *The Computer Journal*, vol. 41, no. 8, pp. 537-546, 1998.

[15] MovieLens 1M Dataset [Internet]. Available: http://grouplens.org/datasets/movielens/1m/.

Artem A. Lenskiy

received his B.Sc. and M.Sc. in Computer Science form Novosibirsk State Technical University, Russia, in 2002 and 2004, respectively. After teaching at the same university for a year, he joined a doctorate course at the University of Ulsan, Korea. He was awarded his Ph.D. in 2010 from the same university. After conducting research as a postdoctoral fellow at the University of Ulsan, he joined Korea University of Technology and Education as an assistant professor in 2011. His research interests include machine learning and applications of self-similar processes to financial time-series, telecommunications, and physiological signals.

Eric Makita

received his B.Sc. and M.Sc. in Computer Science and Engineering from the Institut Supérieur d'Informatique and Ecole Superieure de Technologie et de Management, Dakar, Senegal, in 2005 and 2007, respectively. Currently, he is a PhD candidate at the Korea University of Technology and Education. His research interests include machine learning, big data analysis, and data mining.