Personal Recommender System Based on Agglomerative Clustering together with User-based and Item-based Collaborative Filtering Methods

Ratawan Phantunin*, N. Chirawichitchai
School of Information Technology, Sripatum University, Bangkok, Thailand.
* Corresponding author. Tel.:+ 66-81340-3366; email: aor.tv5@gmail.com

Manuscript submitted February 2, 2020; accepted July 12, 2020.
doi: 10.17706/jsw.15.6.163-171

Abstract: The objective of this study is to develop a Personal Integrated Recommender System. The Recommender System plays an important role and is crucial to our everyday lives in online shopping, meanwhile, it also encounters various problems e.g. scalable data, data sparsity, data accuracy, and having a lot of new users. Therefore, new techniques have been introduced and integrated with the recommender system in order to solve the problems and improve for greater recommender system efficiency. This study, an Agglomerative Clustering together with a User-base and Item-base Collaborative Filtering Method is proposed. By combining the strengths of each method, we can improve the recommender system efficiency and accuracy. The results show that the system being developed generates better values of the area under the curve, precision, normalized discounted cumulative gain, and mean average precision than using only User-based Collaborative Filtering or Item-based Collaborative Filtering alone. Therefore, we can conclude that the Personal Recommender System developed based on Agglomerative Clustering together with User-based and Item-based Collaborative Filtering Method has the ability to increase system efficiency and is applicable. When modern technology arrives in the future, it may reduce the processing time and increase precision

Keywords: Recommender system, agglomerative clustering, user-based collaborative filtering, item-based collaborative filtering.

1. Introduction

The world is moving towards the era of Big Data where we will see the amount of data entering the database is increasing rapidly causing scalability problem, where there are too many data and sparsity problem, where rating per piece of data given by users are not enough to be calculated. This incident generates ideas for the researcher to develop a personal hybrid recommender system to solve the above problems.

The Recommender System [1] is the application of knowledge searching techniques from data in order to use that knowledge for users to make some decisions in order to increase the opportunity for buyers or users to obtain products or information that meets their needs. For instance, Agoda has used a recommender system to recommend accommodations that expect users to pay attention based on the history of other users who have made purchases in the system. In which the system will process data as close as possible to the needs of new users. The collaborative filtering is a highly successful and popular technique for developing the recommender systems [2] for instance, Movie Recommender System, since users are not able to study the details of each and every movie available in the database within a limited time.

Furthermore, by studying the recommender system we found that if the number of users is increasing, it will cause problems with the algorithm e.g. scalability problem which can be solved by using clustering to group users before entering the system to get advice to users faster. Another problem is concerning rating per data piece when there is too much data so the users cannot rate them thoroughly, resulting in insufficient rating for calculating which causing an impact on user satisfaction with the system [3]. While during some research on this matter, we found that there are many studies have suggested ways to solve such problems by using data mining techniques e.g. clustering and classification.

From the above background, the researcher, therefore, developed a personal recommender system based on agglomerative clustering together with user-based and item-based collaborative filtering method. Then used as an experiment for a movie recommender system for users to get effective and accurate recommendations in an efficient time.

2. Theories and Related Research

The recommender system recommends personal information by bringing various information similar to that user with other people who have similar characteristics. It combines and filters for the use of individual suitability. This helps enable users to make quicker decisions and generates more business opportunities due to the ability to recommend products that meet the needs of users the most.

The recommender system has been developed and improved for a long time. Each system has different techniques. The recommender system becomes an important research topic since there is a research under the topic of "Collaborative Filtering". During that time, it was used for business purposes and later has developed in a way that is closer to everyday life by providing a rating or prioritize data to solve problems for users who have never used it before. Sometimes when there is a lot of information, the rating act as an indicator of how much users like using the system. In addition, we are able to add other indicators, including each user's characteristics e.g. age, gender, education, income, etc. The characteristics of each piece of data, including features of movie data e.g. movie titles, movie genres, directors, lead actors, etc. That means that the ability to recommend is replaced by rating with a default system about the user or rating that has been given. The structure of the Recommender System described in Fig. 1 [4], [5].

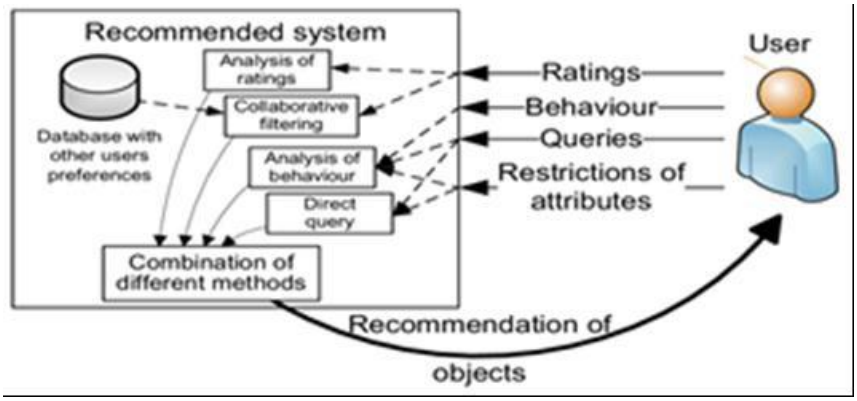

Fig. 1. The structure of the recommender system.

It consists of 1) Data Set, 2) Recommendation Algorithm, and 3) Recommendation List. The data set is the data being taken to analyze to create a recommendation list e.g. Movie Lens with data sets that contains movie data, book item and user profile data, Rating score, etc. This information is being taken to analyze using recommendation algorithm e.g. collaborative filtering algorithm that uses preference rating and personal user data by using various techniques of that algorithm then creating the recommendation list for users in order e.g. hotel recommendations that are suitable for users or recommendation list of clothes that meets

customer preferences, etc. [1]-[5]

In general, the development of the recommender system uses filtering techniques which can be divided into 3 methods: content-based filtering, collaborative filtering, and Hybrid filtering.

2.1. Content-Based Filtering

Content-based filtering is a filtering method used in the recommender system. It is commonly used in information systems that focus primarily on the content according to the data features. This method has an information or content checking process to see any identical or similar data to the user profile, if it's matched, it will bring that data to display as recommendation list to users. If it's not matching, it will not show any recommendation results. Although the information or content may be similar to what the user wants. Therefore, we can conclude that this method calculates only the similarity between the content or data with the user profile by matching the content or data with a user profile to use the information systems based on user interest [4], [5].

2.2. Document Content

Collaborative Filtering is the process of mainly filtering items through the opinions of other users. It will then search for neighboring member groups that have similar preferences with the target group members regardless of the right or wrong of recommendation. This user-based filtering consists of 4 processes: (1) Similarity Computation between 2 users by using 2 similarity calculating methods: Correlation-based and Cosine- based, (2) Neighbor Selection by selecting some users from all users in the system for prediction. There are 2 main techniques used to select neighboring members: Similarity Threshold and Best K Neighbor, (3) Prediction method used to predict user's satisfaction towards one Item by considering the satisfaction and similarities between that item to others which will bring the groups of neighboring members that have already been selected the information to be calculated in order to create further recommendation list and, (4) Creating the Recommendations, by taking the value from the prediction in each item then arrange them into orders. It should start from the items that have the highest prediction value to the item with the lowest prediction value by selecting the number of recommendation items to display. The user can select number of recommendation results they'd like it to display.

Collaborative Filtering is an important process in introducing the "word of mouth" method [3]. Users in the system will assess the liking or disliking of the recommended data which the system will remember and create user history which can be done in 2 ways: obvious and obscure. The obscure way will obscurely refer to what users interested in. This is based on the users' history of products viewed or ordered. As for the obvious, there is an indicator giving ratings to data from what users are familiar with. The recommendation given to the user was made by creating a "Top N" sequence. The algorithm used in collaborative filtering is called a neighborhood-based algorithm. It is one of the many algorithms used with collaborative filtering [6-8] The work-flow of Collaborative Filtering is as shown in Fig. 2.

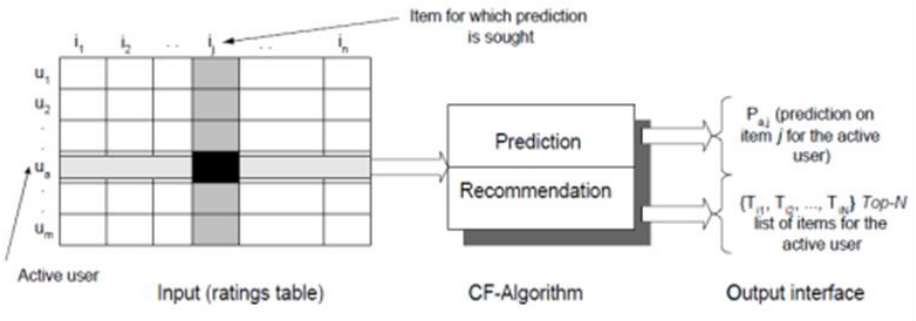

Fig. 2. The collaborative filtering process.

Item recommending by using a lot of similar data in the system or familiar to the user e.g. preferences, movement, and activities of other users in the system. This particular method called the User-based Collaborative Filtering Recommender System. In cases of other characteristics or features of other items in the system, we will call it the Item-based Collaborative Filtering Recommender System. In which using data of other items in the system to be included in the calculation.

2.3. User-Based and Item-Based Collaborative Filtering Method

The process consists of Item-Rating Matrix in searching for groups of similar data pieces. The satisfaction prediction score and output are a result of the predicted satisfaction scores. The item-based collaborative filtering technique based on the foundation of traditional collaborative filtering techniques by considering the relationships between data pieces primarily to search for neighbors. It begins with taking the satisfaction score of the target data piece to compare with every piece of data rating the satisfaction score by users to find a group of data pieces that have the characteristics that give satisfaction scores close or similar to the target data pieces the most. The search begins with a comparison of each piece of data based on other users' opinions who have previously shared satisfaction scores with those two pieces of data which is called co-rated. Then the system will take only the co-rated to calculate and find the similarity between the two pieces of information. The method of calculating similar information - Cosine Similarity is a measuring method of similarities between pieces of information [1-5] which can be explained in equation (1)

$$RSim_{(t,c)} \equiv \frac{\sum_{u \in U} R_{u,t} \times R_{u,c}}{\sqrt{\sum_{u \in U} R_{u,t^2}} \times \sqrt{\sum_{u \in U} R_{u,c^2}}} \tag{1}$$

RSim (*t*, *c*)	means the similarity between data pieces *t* and *c*.
Ru, *t* and Ru, *c*	means the user satisfaction score of u user has for data pieces *t* and *c*
t	means the target piece
C	means a compared data piece
u	means user in the system where *u* = 1,2, ..., *m*

By using the Cosine similarity method to calculate, we'll obtain the similarity of the satisfaction scores that users have given for the data piece in the past.

The process for predicting the satisfaction score is when you found neighboring pieces of data, you can finally bring the data that most similar to the target data to predict the satisfaction scores that users will rate the target data. That is called an item-based collaborative filtering technique based on the weighted sum.

$$P_{u,i} = \frac{\sum_{k \in K}(RSim(t,k) \times R_{u,k})}{\sum_{k \in K}(|RSim(t,k)|)} \tag{2}$$

Pu, *t*	means expected satisfaction score that target user u rates target piece *t*
RSim (*t*, *k*)	means the similarity between the data pieces *t* and *k*
Ru, *k*	means a rating of data item *k* is the score of neighboring pieces of data
K	means the number of neighboring data pieces

The weighted sum equation to predict satisfaction score is a simple equation to predict. It is widely used in collaborative filtering techniques and takes less processing time than the traditional one.

2.4. Agglomerative Clustering

Agglomerative Clustering is a method of grouping. It can be used to group only one or many samples including variables like component analysis. It can also be used with a range of binary data. The agglomerative clustering is a hierarchical grouping method. Those one sample will be grouped with the same one sample by themselves. Those samples with 2 groups that are close to each other will form a hierarchy first. Then consider samples in group 3 to compare the distance whether they are closer to group 1 or 2. If group 3 is closer to group 4 it will combine into a hierarchy 2. It will continue doing this until all the samples are combined into one group. In some hierarchies are nested into subsets of a larger hierarchy which we can see the final will be only one hierarchy. In some hierarchies are set nested into subsets of a larger hierarchy which we can see that in this final hierarchy, there will be only one hierarchy that consists of all sample groups. Although this final hierarchy is not very useful, the structure of the combined hierarchies may indicate the number of major hierarchies that can be separated from other hierarchies and be used further to determine the appropriate hierarchy [9], [10].

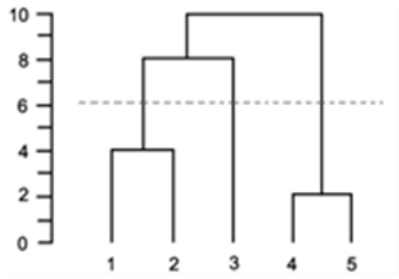

Fig. 3. Agglomerative clustering.

3. Theories and Related Research

3.1. Research Methodology

The system presented in this research focuses on the development of a hybrid recommender system by agglomerative clustering together with the user-based and item-based collaborative filtering method. The method uses a lot of similar data in the system or familiar to the user e.g. preferences, movement, and activities of other users in the system. It is called User-based Collaborative Filtering Recommender System. In cases of other characteristics or features of other items in the system, we will call it Item-based Collaborative Filtering Recommender System. This type of recommender system typically has 3 major processing steps as follows: 1) create user history or data to be used as the system basis, 2) select a co-rated item or co-user rate that is close or similar to a predetermined amount by comparison of user history or data used as the system basis which may use same values as Cosine Similarity [1]-[5] to calculate how well users and data are matched.

3.2. Data Collection

Data in this research experiment is a basic set for creating a system database and for grouping users. The data is taken from the MovieLens Project. The data set includes 1. Movie Rating data set of 1,000,000 Records, 2. User data set consists of gender, age, occupation, and postcode, and 3. The movie data set e.g. action, adventure, comedy, drama, etc. The data used in the test is a sample of the actual data collected. The users have accessed to provide information in the MovieLens Project about movies they have seen. From studying problems and collecting data, the developer has designed the research algorithm as shown in Fig. 4.

3.3. System Testing and Evaluation

Research on the development of a personal recommender system based on Agglomerative Clustering

together with user-based and item-based collaborative filtering method considers the efficiency of the model by measuring from AUC: the area under the curve, prec: precision, NDCG: normalized discounted cumulative gain, and MAP: mean average precision, which will be explained as follows, [11], [12].

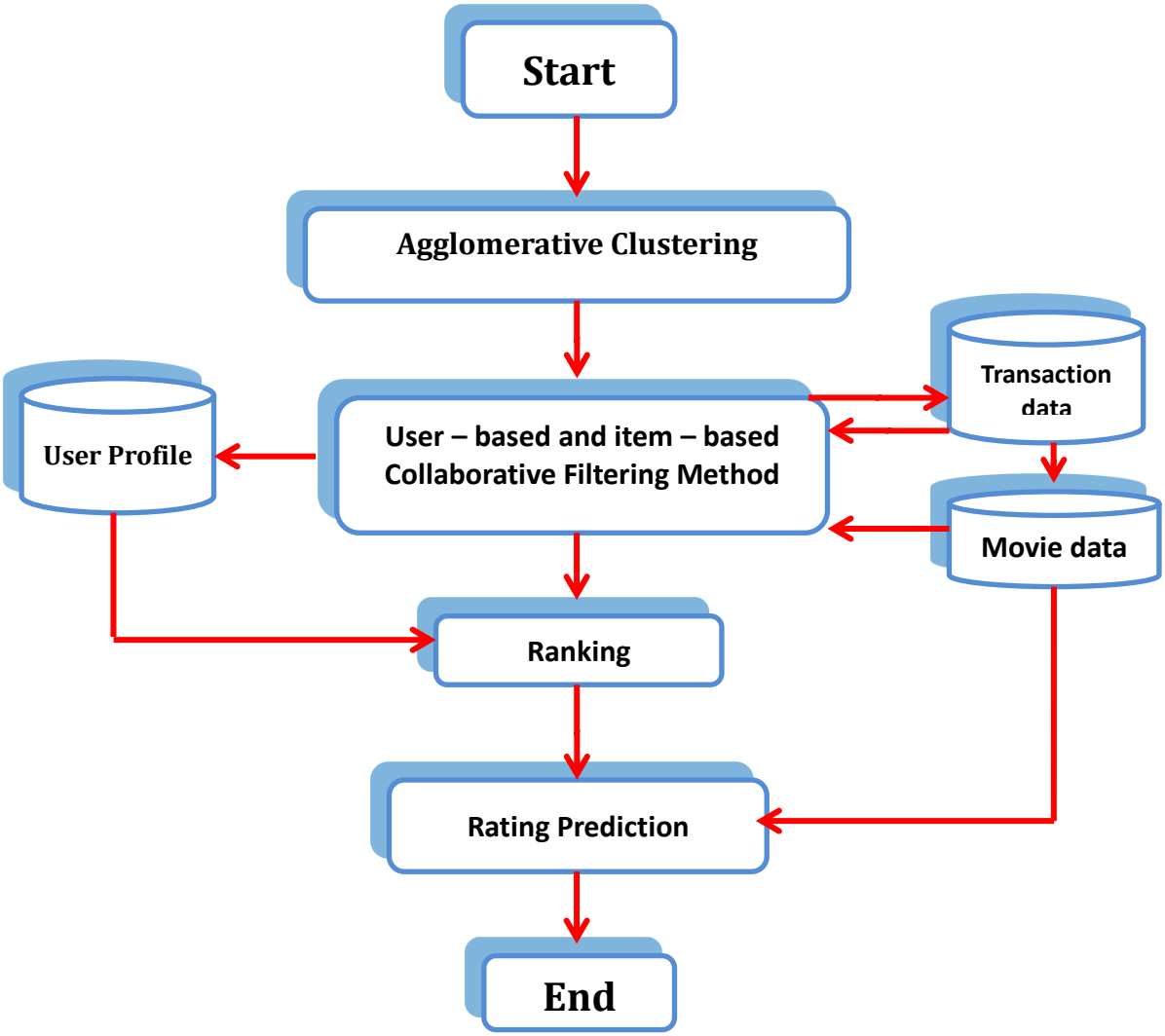

Fig. 4. Process of the personal recommender system based on agglomerative clustering together with user-based and item-based collaborative filtering method.

Table 1. The Search Results

	Retrieved	Not Retrieved
Relevant	*tp*	*fn*
Not Relevant	*fp*	*tn*

Precision = $tp / (tp + fp)$

1) Area Under the Curve (AUC) is another alternative for measuring the performance of the model. The x-axis is the true positive rate and the y-axis is the false positive rate. The AUC measurement starts from 0 to 1. 0 means the model is underperforming and 1 means the model has the highest efficiency. This measurement result will be easily understood.
2) Precision (prec) states that it is a measure of the system precision to find relevant documents correctly.

The results are as shown in Table 1.

Table 2. Experimental Results in the Case of Using Agglomerative Clustering Together with User-Based Collaborative Filtering and Item-Based Collaborative Filtering Being Classified into a Table

	User k-NN	Item k-NN	Model Combiner
AUC	0.913	0.912	**0.928**
Prec@5	0.398	0.272	**0.415**
Prec@10	0.338	0.252	**0.358**
Prec@15	0.301	0.234	**0.320**
NDCG	0.591	0.530	**0.604**
MAP	0.223	0.167	**0.241**

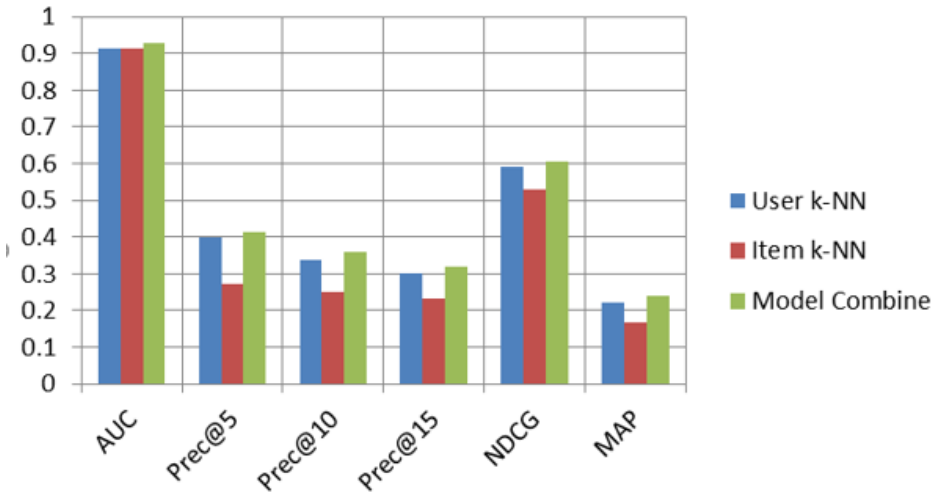

Fig. 5. Experimental Results of using agglomerative clustering with user-based collaborative filtering and item-based collaborative filtering.

3) Normalized discounted cumulative gain (NDCG) is a theory used in the evaluation of search engines result. The level of relevance of the document including considering the ranking of search results. The DCG is a measurement of the document suitability interested in the position or order of the document. The grade points received are cumulative from the top of the search results to the bottom of the search results list. The scores will decrease when satisfaction results are ranked low [12] the NDCG equation is as follows;

$$NDCG_q = M_q \sum_{j=1}^{k} \frac{(2^{r(j)}-1)}{log(1+j)} \quad (3)$$

where k is the level or criterion used, r (j) is the relative value of the document that is evaluated by the user, Mq is the constant value of the ordering completeness with the highest value at 1. NDCG will reward the

related documents that appear in the ranking of search results and punish unrelated documents by reducing NDCG scores.

4) Mean Average Precision (MAP) is the average accuracy of many relevant key-words and ranking with the equation as follows;

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \quad (4)$$

Q means the number of words searched

4. Results

After the experiment, we found that the Model Combiner method using agglomerative clustering together with user-based and item-based collaborative filtering, the Area Under the Curve (AUC) was at 0.928, precision (prec) k@5 was at 0.415, k@10 was at 0.358, and k@15 was at 0.320, normalized discounted cumulative gain (NDCG) was at 0.604, mean average precision (MAP) was at 0.241, all of which were significantly higher than those from user-based collaborative filtering and item-based collaborative filtering.

5. Discussion

This research has developed a recommender system based on agglomerative clustering together with user-based and item-based collaborative filtering methods, by combining the strengths of each method. This result in the recommender system to be more efficient in all indicators consists of the Area Under the Curve (AUC) is at 0.928, precision (prec) k@5 is at 0.415, k@10 is at 0.358, and k@15 is at 0.320, normalized discounted cumulative gain (NDCG) is at 0.604, mean average precision (MAP) is at 0.241. Therefore, it can be concluded that the personal recommender system based on user-based and item-based collaborative filtering methods helps supporting users' decision making more precisely than the personal recommender system that uses only user-based collaborative filtering alone. It also helps in solving scalability and sparsity problems and is also applicable.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Ratawan Phantunin analyzed the data and wrote the paper; Nivet Chirawichitchai conducted the research; all authors had approved the final version.

Reference

[1] Adomavicius G., & Tuzhilin A. (2004). Recommendation technologies: Survey of current methods and possible extensions. *Collections of Information Systems Working Papers*. Stern School of Business, New York University. Report No. IS-03-06

[2] Sarwar, B. M., Karypis, G., Konstan, J. A., & Riedl, J. T. (2000). Analysis of recommendation algorithms for e-commerce. *Proceedings of the 2nd ACM. Conference on Electronic Commerce*. 67-158.

[3] Haruechaiyasak, & Mei, L. S. (2005). Collaborative filtering by mining association rules from user access sequences. *Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration* (pp. 128-135).

[4] Felfernig, A., Friedrich, G., & Schmidt-Thieme, L. (2007). Recommender systems. *IEEE Intelligent Systems Special Issue on Recommender Systems*. IEEE Computer Society.

[5] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of thestate-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering, 17(6)*, 734–749.
[6] Nutcha, R. M. (2009). Recommender system using pseudo rating and multidimensional data. Master's Thesis, Faculty of Science, Chulalongkorn University.
[7] Schafer, J., *et al*. (1999). Recommender systems in e-commerce. *E-commerce 99*. Colorado: Denver.
[8] Chirawichitchai, N., (2015). Developing term weighting scheme based on term occurrence ratio for sentiment analysis. *Information Science and Applications*. Springer, 737–744.
[9] Refat, A. (2017). Agglomerative hierarchical clustering: An introduction to essentials. Proximity Coefficients and Creation of a Vector-Distance Matrix and (2) Construction of the Hierarchical Tree and a Selection of Methods.
[10] Janpla, S., & Wanapiron, P. (2018). System framework for an intelligent question bank and examination system. *International Journal of Machine Learning and Computing, 8(5)*.
[11] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters,* 861–74.
[12] Smoot, B. J., Wong, J. F., & Dodd, M. J. (2011). Comparison of diagnostic accuracy of clinical measures of breast cancer–related lymphedema: Area under the curve. *Archives of Physical Medicine and Rehabilitation*, 603-610.
[13] Jarvelin, K., & Kekalainen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS), 20(4)*, 422–446.

Ratawan Phantunin was born in Bangkok Thailand in 1978. In 2002, She obtained a bachelor's degree in international business from the University of the Thai Chamber of Commerce, Thailand. In 2005, the author obtained a master of education in educational technology from the Faculty of Education, Kasetsart University and now she is a Ph.D. candidate at School of Information Technology, Sripatum University, Thailand.

The author's major fields of study are data mining and big Data. Since 2016, she has been working as a senior policy and planning officer at the Office of the National Broadcasting and Telecommunication Commission,Thailand.

Nivet Chirawichitchai received his Ph.D degree in information technology from King Mongkut's Institute of Technology North Bangkok, Thailand. He currently has the rank of director, a master of science program in information technology, School of Information Technology, Sripatum University, Thailand. His main research interests are in the field of machine learning, data mining and big data analytics and several published articles in these fields. He is also an assistant professor at graduate school in Information Technology Department, Sripatum University.