Chapter 1

Introduction

1.1 Background and Motivation

There has been an explosion, in recent years, on the number of web sites available on the internet. This can result in the user being inundated with a great number of web sites to view, highlighting the subject matter and providing reams of information, most of which will be irrelevant for the users' needs. To solve this problem, a Recommendation System (RS) has been developed which filters the web sites on the internet, only recommends to the user those that will be most appropriate to view. The methods of the RS can be broadly categorized as: (1) content-based [1, 2], (2) collaborative filtering [1-6], (3) feature-based [7, 8], and (4) demographic-based algorithms [9, 10]. Collaborative filtering algorithm is widely applied in the area of the RS [1-6, 11, 12]. It may be classified into two methods, i.e., user-based and item-based algorithms. In the user-based collaborative filtering algorithm, the system generates top-*N* recommendation based on similarity among users. In the item-based collaborative filtering algorithm, the system generates top-*N* recommendation based on similarity among items. For example in commercial applications, Amazon.com and CD-Now.com have developed their recommendation systems using an item-based collaborative filtering method [5, 11].

Although, there are many recommending algorithms developed to improve the performance of the recommendation systems. The recommendation system still needs

to overcome the cold-start and sparsity problems [13]. The cold-start problem occurs when the recommendation systems does not have enough information about a new user and/or new item as shown in Figure 1.1. The sparsity problem occurs when the frequency of the purchased items is too small as shown in Figure 1.2. The black points are represented the existing information between users and items. The white points are not have the information between users and items.

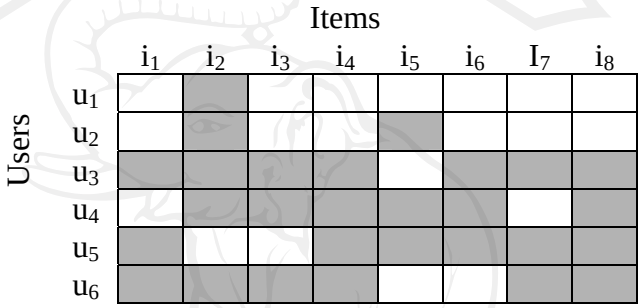

Figure 1.1 Example of the cold-start problem.

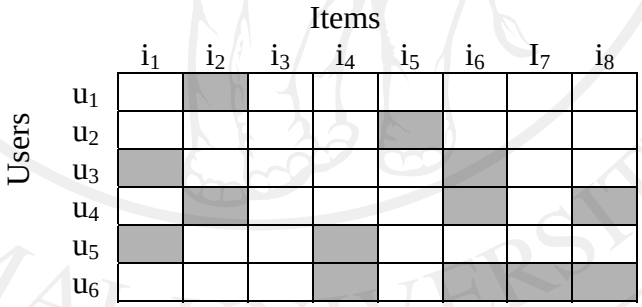

Figure 1.2 Example of the data sparsity problem.

There are many algorithms used to address the sparsity problem such as combining collaborative filtering with personal agents [13], using the fuzzy-based method [14], and combining collaborative and content-based filtering [2]. For the cold-start problem, this issue can be divided into problems with new users and new items. To solve the new user problem, the fuzzy-based, item-based collaborative filtering, feature-based methods were used [5, 7, 14, 15]. Furthermore, the fuzzy-based and feature-based methods were used to solve the new product problem [7, 14,

15]. However, all the mentioned methods have limitations. In the fuzzy-based method, the performance of the method depends on the expert who creates the fuzzy rule-base. In combining between collaborative filtering with personal agent's method, the personal information is difficult to collect. When using the combining collaborative and content-based filtering method and the feature-based method, items have the same attributes such as book, movie, and song. Above all, the algorithms still have problems when the cold-start and sparsity problems occur together. These problems lead to low performance in recommendations in the real-world data sets. So, the objectives of this research are to design the image-based clustering method for recommendation system to solve the cold-start and sparsity problems, and to implement a proposed method for testing the performance of the recommendation system in the real-world data sets.

1.2 Literature Review

Salter and Antonopoulos [2] developed the recommendation system by combining collaborative and content-based filtering technique to improve the accuracy of prediction rating. There are two processes for computing a rating of an active user. The first process was to calculate the rating of each film by using the collaborative filtering technique. The rating results input from the first process are the input of the second process. The rating of each film was re-calculated by using the content-based filtering technique based on the information of the films and the rating from the first process. The proposed technique was compared with a collaborative filtering, a content-based filtering, and a collaboration-via-content filtering techniques. The results showed that this method produces less mean absolute error than traditional techniques, such as the collaborative filtering and the content-based

filtering techniques. However, the items that used for the recommendation system must have the same attributes.

Zhang and Chang [3] solved the scalability problem for the collaborative filtering algorithm. The genetic algorithm was applied to solve the scalability problem by clustering users into groups. The chromosomes of the genetic algorithm were designed with the number of clusters. The fitness function of the genetic algorithm was computed with the cluster's intra-similarity. The proposed method was evaluated by using MovieLens data set and mean absolute error. In the experimental result, the proposed method was compared with the memory-based collaborative filtering and *k*-means clustering based on collaborative filtering. The experimental result shows that the proposed method is the best. The disadvantage of this method is the number of clusters extremely hard to assign.

Gong [4] solved the sparsity problem in the personalized recommendation by joining case-based reasoning and item-based collaborative filtering. Firstly, the case-based reasoning was used to fill empty ratings in the user-item matrix. Then, the item-based collaborative filtering was applied to predict the rating of an active user. For experimental result, this method was compared with the user-based collaborative filtering and reported by using ROC-4. It is found that this method increased the performance of the system. The personal information is the limitation of this method because it is the private information.

Linden et al. [5] developed the item-to-item collaborative filtering algorithm for Amazon.com shopping cart recommendations to address the scalability problem. Rather than building a similarity-items matrix by computing similarity between items,

the proposed algorithm built a similarity-items matrix by iterating through all item pairs and computing a similarity metric for each pair. In the experiment, the cosine measure was selected to compute similarity between items and MovieLens data set to evaluate performance. The experimental results show that computing similarity to build similarity-items matrix depended only on the number of items the customer purchased. However, the cold-start and the sparsity problems still exist.

Mema and Shikha [6] proposed the recommendation system by using the memetic collaborative filtering. This method is similar to the recommendation system using the GA *k*-means proposed by Kim and Ahn [16]. The difference is the genetic and memetic algorithms. This method used the memetic algorithm to improve the cluster performance of the *k*-means. The recommendation process used the derived clusters to predict and recommend items to the active user based on the user's preference rate. The MovieLens data set with 730 users was used to evaluate the performance of the recommendation system. This method was compared with the traditional collaborative filtering by using the F-measure. The results of all the experiments show that the performance of the proposed method is better than the performance of the traditional method. However, limitations of this method are that the number of clusters is hard to define by the user who created the system and it is hard to generate the items to an active user based on rating under cold-start and sparsity problems.

Weng and Liu [7] proposed the feature-based recommendation system. To address the brand new items that often lack rating was the point of this research. There were two procedures in the recommendation system. The first procedure was

recommendation based on a customer profile module. Second procedure was recommendation based on a customer cluster profile module. In the first procedure, the customer profile was created from item profile module, creating an item profile module. The features of item were defined into 1 (the feature value) and 0 (otherwise). Then, the system calculates similarity between the profiles of the active customer with the group of items. In the second procedure, the customer profiles were clustered by using two-stage clustering techniques. For two-stage clustering techniques, the *k* number of clusters was discovered using the self organizing maps (SOM). *k* was assigned to the centroid of *k*-means clustering technique. The result of each cluster analyzed was then integrated with the first procedure for recommendation. Precision, recall, and F1 were also used to evaluate this recommendation system. The feature of items is the limitations of this method because the features of each item must have the same.

Adomavicius and Kwon [8] proposed a new technique for multi criteria rating systems to recommend movies. To improve the performance of recommendation system, the single rating was extended to five rating. Memory based and model based methods were used to compute similarity between users on a multi criteria rating. Thus the user-to-user collaborative filtering method used to produce top-*N* recommendation to the active user. They collected a data set from 155 users and 50 movies, from Yahoo! movies to test performance. The experiment found that this method produced a higher accuracy of recommendations than traditional methods. However, the system requires additional information from users to create a more accuracy for recommendation.

Cho and Li [14] proposed the fuzzy-based model to recommend the items to the users who have less information. The process of this model consisted of creating the fuzzy rule base; getting the current information of active customers, and recommending items to the active customers. A data set, 128 laptops of different brands, was collected from Amazon.com. Recall, precision, and F1 were evaluated by the recommendation system with seven customers. The average of Recall, precision, and F1 were 83.82%, 87.57%, and 85.39%, respectively. However, the performance of the recommendation depended on the experience of experts.

Debnath et al. [15] proposed the feature weighting in content based recommendation. The features of the movie were considered to improve the recommendation. The features that used in the movie recommendation are release, type, rating, vote, director, writer, genre, keyword, cast, country, language, color, company. The feature weights were estimated from a social network graph of items. The optimal feature weights were included into a similarity measure between items to recommend. The movie data set was downloaded from the IMBD to evaluate the performance of the recommendation. The proposed method was compared to the pure content based method. The recall results show that the proposed method achieves the average recall of 0.29. Where as, the pure content based method achieves the average recall of 0.24. However, only the items which had the same feature could be used in this method

Kim and Ahn [16] proposed the GA *k*-means clustering technique. The problem of the *k*-means clustering technique is in the selection of the initial seeds. Normally, the initial seeds fall into local optimization. To solve the initial seed

problem a genetic algorithm was used to discover the initial seeds within the *k*-means clustering technique. A chromosome of the GA consisted of binary values and five clusters. Calculating the intra-class inertia was defined as the fitness function. From an online diet portal site in Korea, a data set was evaluated using the GA *k*-means clustering technique. The proposed method compared the results of the method to a *k*-means algorithm and self-organizing maps. In intra-class inertia of each clustering, GA *k*-means is the best among the proposed methods. The number of clusters is the limitations of this method because the number is hard to define by users.

JingHui et al. [17] investigated two kinds of methods used in the recommendation system. Both collaborative filtering and clustering methods were analyzed. They found that the collaborative filtering methods consisted of the content-based filtering and item-based filtering for using recommendation and prediction. In addition, the clustering technique was applied to improve the recommendation quality. The different content of item is the limitation of this method.

Park and Chang [18] proposed the new customer profile model to improve the quality of recommendation. The customer profile model was developed based on individual and group behavior information. There were two processes for creating the customer profile: (1) creating an item profile and (2) creating the customer profile from the item profile. In the customer profile module, it included the customer transaction histories (clicks, basket insertions, purchases, and interest fields). It also included the information of group interests. To produce items to an active customer, the recommendation system computed the similarity between the active customer and each item group by using the Euclidean distance. The performance of the

recommendation system was compared with two customer profile modules that are individual purchasing information and individual behavior information. Mean absolute error, precision, recall, and F1 were measured to investigate the performance of the recommendation system. The proposed method was able to effectively recommend items to the active user. However, only the items which had the same feature could be used in this method.

Lu and Li [19] developed the collaborative filtering recommendation system in the process of searching for nearest neighbors. To calculate the similarity value between the target customer and the other customers, the genetic algorithm was used to optimize the weight value. The proposed technique calculated the similarity by using the information of the customers instead of voting score. The MovieLens data set was used to evaluate the system. In this method, it is hard to calculate the similarity between users under cold-start and sparsity problems.

Gao and Li [20] proposed the hybrid recommendation model. To increase accuracy of the recommendation system, the genetic algorithm was used to find the weight vectors for hybrid recommendation system. The result of the optimal weights from the genetic algorithm was used to calculate in hybrid recommendation model. Mean absolute error is used to evaluate the quality of the recommendation. The proposed method was compared with the content-based recommendation and the collaborative filtering recommendation. The mean absolute error of the proposed method was rather less than another. On the other hand, it was time consuming for calculating each module of recommendation system. The increasing customers and items are also problem of this method.

Zanker et al. [21] compared the recommendation strategies. There were three recommendation methods which were compared: (1) the top-*N* recommendation based on sales records, (2) content-based algorithm retrieves, (3) hybrid between the content-base algorithm and the collaborative filtering algorithm. Four different data sets were collected for testing the recommendation system. Recall, user coverage, and catalog coverage were used to evaluate the performance of the recommendation systems. The experimental results show that the hybrid between the content-base algorithm and the collaborative filtering algorithm has the best performance. The different content of item is the limitations because the content must have the same.

Poonam [22] proposed the comparison results of the cryptanalysis of simplified data encryption standard problems (SDES) using the genetic and memetic algorithms. The main objective of this research is to use the genetic and memetic algorithms to investigate the performance of SDES encryption algorithm. The input and output of the genetic and memetic algorithms is an 8-bit block of plaintext. The experiment results show that the accuracy of the memetic algorithm is better than that of the genetic algorithm. Comparing the running time of the memetic and genetic algorithms, the results show that the memetic algorithm used more running time than the genetic algorithm. Because the memetic algorithm is the extension of the genetic algorithm, the running time of the memetic algorithm is higher than the genetic algorithm. The disadvantage of this method is the process time because the memetic algorithm is the extension of the genetic algorithm. The process time of the memetic algorithm uses more time than the genetic algorithm.

Zexuan, Sen, and Zhen [23] proposed the feature selection using the memetic algorithm. In this method, the memetic algorithm was applied to select the property features in the microarray and hyperspectral imagery data sets. The features of the data sets were designed into zero and one. This method integrated the add and delete operators in the local search of the memetic algorithm. There are three proposed methods, i.e., Filter Ranking, Approximate Markove Blanket, and Affinity Aropagation, compared with the feature selection methods based on genetic algorithm, i.e., Filter methods and Fast Correlation-based Filter. The experiment results show that the feature selection method based on the memetic algorithm provided better performance than the traditional methods.

In the literature review, there are many methods for the RS such as the feature-based, collaborative filtering, content-based, and demographic methods. In the content-based method, Salter and Antonopoulos [2], Gong [4], JingHui et al., Debnath et al. [15], Gao and Li [20], Zanker et al. [21], and Weng and Liu [7] applied this method for the RS. The feature-based method was applied by Weng and Liu [7] to the RS. In the collaborative filtering method, Salter and Antonopoulos [2], Zang and Chang [3], Adomovicius and Kwon [8], JingHui et al. [17], Lu and Li [19], Gao and Li [20], and Zanker et al. [21] applied this method for improving the performance of the RS. To solve the cold-start problem in the RS, Wang and Liu [7], Cho and Li [14], and Park and Chang [18] used the feature-based, fuzzy-based, and customer profile, respectively. To solve the sparsity problem in the RS, Gong [4] applied the hybrid between the case-based reasoning and item-based methods whereas Cho and Li [14] used the fuzzy-based method. The GA also used to improve the performance of the RS. Kim and Ahn [16] and Poonam [22] used the GA and *k*-mean for the RS. The

memetic algorithm also used to improve the performance in the RS. Mema and Shikha [6] applied the MA to improve the performance of the RS. However, the limitation of the content-based methods is that the content of items must have the same. The feature-based method that item must have the same attributes is the limitation. Creating the fuzzy rule-based by the experts is the limitation of the fuzzy-based method.

1.3 Purposes of the Study

1.3.1 To find new clustering methods for the recommendation systems.

1.3.2 To illustrate that the new clustering methods can reveal the relationship between users and items.

1.3.3 To apply the proposed methods in the recommendation systems with real-world data sets.

1.3.4 To improve the performance of recommendation systems under the cold-start and sparsity problems.

1.3.5 To compare the performance of the proposed method with common methods, i.e., the *k*-means clustering method, frequency-based method, user-based collaborative filtering method, and item-based collaborative filtering method.

1.4 Research Scope and Method

1.4.1 The problems in recommendation systems considered in this research include the cold-start and sparsity problems.

1.4.2 The proposed methods are compared with four common methods, i.e., the *k*-means clustering method, frequency-based method, user-based

collaborative filtering method, and item-based collaborative filtering method.

1.4.3 The F-measures including precision, recall, and F1 are used as the evaluation measures in this research.

1.4.4 The real-world data sets used in this research are the KDD-CUP 2000 collected by Blue Martini Software, Inc., the TTS collected from Thaiherbs-Thaimassage.com, Thailand, the ECR collected from the University of California on the UCI machine learning repository, the RCM collected from the Department of Computer Science, National Center for Research and Technological Development in Mexico, and the MovieLens collected by the GroupLens Research Project the University of Minnesota.

1.5 Education/Application Advantages

1.5.1 Obtain new clustering methods for clustering customers and items.

1.5.2 Obtain new methods for the recommendation systems under the cold-start and sparsity problems.

1.6 Research Methodologies

1.6.1 Study the fundamental algorithms of clustering and top-*N* recommendation system.

1.6.2 Study the problems of top-*N* recommendation system in real-world data sets.

1.6.3 Design new methods for top-*N* recommendation system under the cold start and sparsity problems.

1.6.4 Implement the methods for top-*N* recommendation systems under the cold-start and sparsity problems.

1.6.5 Evaluate the performance of the proposed methods by comparing with the common methods.

1.7 Organization of Dissertation

The thesis is divided into five chapters. Chapter 1 begins with the introduction. Chapter 2 reviews the clustering, top-*N* recommendation system, the genetic, and memetic algorithms. Chapter 3 describes the research designs and the proposed method. Chapter 4 describes the experimental results of the proposed method on synthetic data sets and real data sets. Finally, conclusions are drawn in Chapter 5.