Detection of Illegal Players in MMORPG by Classification Algorithms

Zhang, Zhongqiang
Graduate School of Information Science and Electrical Engineering, Kyushu University

Anada, Hiroaki
Institute of Systems, Information Technologies and Nanotechnologies

Kawamoto, Junpei
Institute of Systems, Information Technologies and Nanotechnologies | Faculty of Information Science and Electrical Engineering, Kyushu University

Sakurai, Kouichi
Institute of Systems, Information Technologies and Nanotechnologies | Faculty of Information Science and Electrical Engineering, Kyushu University

出版情報：Proceedings of the 29th IEEE International Conference on Advanced Information Networking and Applications, 2015. IEEE Computer Society
バージョン：
権利関係：

Detection of Illegal Players in Massively Multiplayer Online Role Playing Game by Classification Algorithms

Zhongqiang Zhang*†, Hiroaki Anada†, Junpei Kawamoto†‡, Kouichi Sakurai†‡
* Graduate School of Information Science and Electrical Engineering, Kyushu University
744 Motooka, Nishi-ku, Fukuoka, JAPAN
Email:zhangzq@itslab.inf.kyushu-u.ac.jp
†Institute of Systems, Information Technologies and Nanotechnologies
Fukuoka SRP Center Building 7F
2-1-22 Momochihama,Sawara-ku,Fukuoka, JAPAN
Email:anada@isit.or.jp
‡Faculty of Information Science and Electrical Engineering, Kyushu University
744 Motooka, Nishi-ku, Fukuoka, JAPAN
{kawamoto, sakurai}@inf.kyushu-u.ac.jp

***Abstract*— Online games have become one of the most popular games in recent years. However, fraud such as real money trading and the use of game bot, has also increased accordingly. In order to maintain a balance in the virtual world, the operators of online games have taken a stern response to the players who conduct fraud. In this study, we have sorted out players' behaviors based on players' game playing time in order to support and find potentially illegal players in the MMORPG. In this paper, we added a topic model to the experiment and used k-means as a major tool to classify the players in the World of Warcraft Avatar History Dataset and find potentially illegal players.**

Keywords— Online game; fraud; topic model; k-means; classify

I. INTRODUCTION

Online game is a kind of computer game that many participants can play game simultaneously over the Internet, and it has become one of the most popular games in recent years. Especially, in the massive multiplayer online role-playing games (MMORPG), the notion of "clear" continues to expand and there are various ways to enjoy the game such as becoming stronger by level-improvement, collecting rare items, and communicating with other players among the online participants. It is even possible to defeat a strong enemy by cooperating with others. However, fraud such as real money trading and the use of game bot, also increased accordingly. Enemy can be beaten down automatically by the bot, and it is possible to acquire a large number of items that were dropped from the enemy. As a result, the balance of the game will be broken. There are many kinds of bots with different functions, namely some bots that are designed for raising the game level, earning for the currency, or even making a quest. It is difficult to detect the bots, for that they are designed to simulate the behavior of a human player and usually follow the rules of the game.

In order to maintain the balance in a virtual world, the operator of online games has taken a stern response to the players who perform a fraud. For example, players using a bot or doing a real money trading will be prohibited from the game. However, this inspection process requires the amount of labor of humans.

In this paper, we did a classification of player's behaviors in order to do a support to find the illegal players in MMORPG. Specifically, we used a feature vector to represent a player's actions, and then we used k-means algorithm to classify players into several groups. World of Warcraft Avatar History (WoWAH) Dataset[1] [1], which is published in the internet, was used in this paper. We first did an experiment by using k-means algorithm, then we added a topic model to the experiment and obtained the results.

K-means algorithm is shown as follows. First, we take out the Avatar ID of all players from the dataset. And then we need to calculate the time at each level of the players. Since there is only sampling time in the dataset, players' game playing time can be calculated on the basis of the sampling time. Next, we used k-means algorithm to divide the players into several groups. The performance of k-means algorithm is affected by the method of calculating the similarity and the number of the group that it is to be classified in this step. In this paper, we selected some number of groups to do the experiments, and compared the results.

We used Latent Dirichlet Allocation (LDA) algorithm in this paper for the topic model, in which the vectors of players were regarded as probability model under some topic. In this method, we generated a probability model by the LDA algorithm. And then, we divided players into several groups through k-means algorithm. Finally, we found that it was possible to detect the irregular players whose actions were

1http://mmnet.iis.sinica.edu.tw/dl/wowah/

TABLE I. EXAMPLE AVATAR OBSERVATION RECORDS

Query Time	Seq.#	Avatar ID	Guild	Level	Race	Class	Zone
01/01/06 23:59:39	1	467		1	Orc	Warrior	Orgrimmar
01/01/06 23:59:39	1	921	19	1	Orc	Shaman	Orgrimmar
01/02/06 00:03:31	45	1367	8	60	Undead	Warrior	Arashi Mountain

TABLE II. FIELD DESCRIPTION

Start date	2006-01-01
End date	2009-01-10
Duration	1,107 days
Sampling rate	144 samples per day
# of samples	159,408
# of missing samples	21,324
# of avatars	91,065
# of sessions	667,032

different from the average players by comparing the characteristics of the categorized players.

Section II reviews the related works. Section III describes the World of Warcraft Avatar History (WoWAH) dataset. The remainder of this paper is organized as follows. Section II reviews the related works. Section III describes the World of Warcraft Avatar History (WoWAH) dataset. In section IV, we discuss the balance of the game world. Frauds such as real money trading and game bots can break the balance of games in cases. Section V describes the k-means algorithm, and we propose our design of the featured vector for WoWAH dataset. Section VI describes the LDA algorithm, a kind of generative probabilistic model which can transform articles into a topic model. In section VII, we evaluate the performance of the proposed schemes and discuss the players' characteristic and activities. Our conclusion will be in Section VIII.

II. RELATED WORK

While frauds are regarded as a crucial challenge to the design of online game, a great deal of effort has been devoted to cheat the prevention schemes.

Chen et al. took the moving locus of the avatar in FPS game, and used machine learning to detect the bot [2]. In [3], Chen et al. proposed a scheme to solve the identity theft and sharing the problems by exploiting players' game-play activity patterns, which took the idle periods between successive moves of a player-controlled character to characterize player's game-play characteristics. And in [4], Chen et al. also proposed two ensemble schemes to identify game bots by studying the traffic of human players and mainstream game bots under different network settings. Thawonmas et al. proposed to do a classification of game players based on their movement trails using self-organizing map [5].

Fujita et al. proposed a method for detecting real money trading by taking the volume of the trading between players in the Massively Multiplayer Online Role-Playing Game (MMORPG) [6]. Hiroshi, et al proposed several statistics derived from real log data of a MMORPG to differentiate the RMT (Real Money Trade) players from general players [7].

Although many works has been done to prevent from cheating, frauds may still be active in online games. It was announced on the Internet that three suspects were arrested for using the cheat tool in online game "Sudden Attack"[2].

III. OVERVIEW OF THE DATASET

From the perspective of a game designing, one of the most important factors is the behaviors of players while designing a game. It is a good point to do a research of the actions of game players by investigating the game playing time of the players. The concept of the game playing time is that it is applicable to any kinds of games, by which it is possible to make a model of the system load and the effect of the system and network's QoS (Quality of Service). Furthermore, the operating companies of the games can perform prediction on the activities of the game players in some particular games through the research of the behaviors of players.

In this paper, we used the dataset of WoWAH as a study case to do the experiments. "World of Warcraft "is one of the most popular MMORPGs, which obtains over 12 million active plays [8].

Table I shows the dataset sample of WoWAH. The 8 fields are query time, query sequence number, avatar ID, guild, level, race, class, zone. The meanings and valid values of the fields are detailed in Table I. Each element stores the information about an avatar observed during the sampling period; thus, the number of elements is equal to the number of avatars online in that sampling interval.

Table II shows the summary of the dataset. the dataset from January 2006 to January 2009 was sampled for every 10 minutes. And there are 91065 of avatar ID recorded in the dataset. Moreover, for there are 21,324 of missing samples of the dataset, it will cause some errors on the results.

In World of Warcraft, players need to pay $15 [9] to purchase a subscription for the service of the game every month. There are two kinds of categories, items and currency, that can be collected by players in the game. Currency takes a form of coins, for example, with one hundred bronze coins to one silver coin and one hundred silver coins to one gold coin [10].

In order to get more items and currency or to raise level of the avatar, some players may do illegal actions. For example, they may do real money trading or use a game bot. These kinds of activities may break the balance of game in cases, and it can

2 http://nlab.itmedia.co.jp/nl/articles/1406/25/news119.html

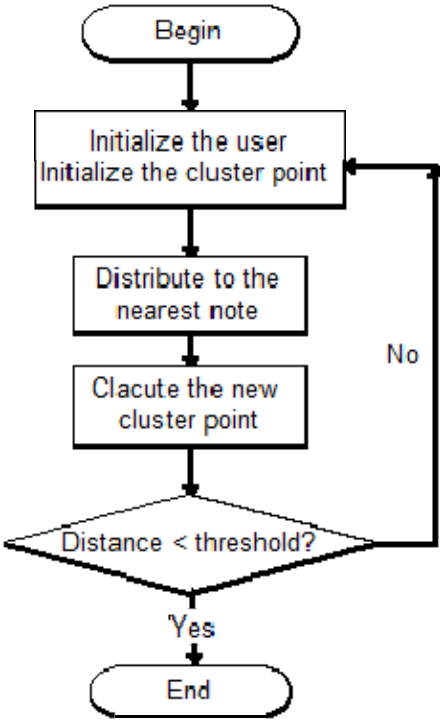


Fig. 1. flowchart of the k-means.

also cause trouble among game players and bring trouble to the company operating the of the game.

IV. THE BALANCE OF GAME

The frauds such as real money trading and game bots can break the balance of games in cases.

A. Game Bot

One of the MMORPG's(Massively Multiplayer Online Role Playing Game) greatest challenges is the increasing use of game bots, that is, auto-playing game clients. The etymology of bot is robot, and it is a kind of program that can play the game intelligently and is often used in the FPS(First Person Shooter) and MMORPG. In order to achieve a special goal, bots are usually designed to repeat some specific action. Player can use a game bot to knock down the enemy automatically, as a results the player can get a large quantity of items. Players can get a large amount of items and currency automatically in the game in this way, and it means that the price of items in the game world will increase rapidly. As a result, it may bring the trouble to the general players. It is possible to play game 24 hours a day with a game bot, and it may break the balance of the game in case. Actually, there are many kinds of bots with different functions, such as bots designed for raising the level, earning the currency, or doing a quest.

B. Real Money Trade

Real Money Trade is the activity that players do some buying and selling in the real world use items and currency that obtained in the game. The items that are hard to be found are called rare item in the online game, they may also be traded at

Algorithm 1 k-means algorithm

```
Input: players: the object of the players
Output: marked players: the players is marked by labels
players = readObject ();
ClusterPoint= randomSelect ();
for distance < threshold do
   for player in the players do
      for point in the ClusterPoint do
      find the nearest point;
      mark the player;
      end for
   end for
   ClusterPoint = getNewPoint();
   calculate distance;
end for
return players
```

Fig. 2. k-means algorithm.

several millions of yen in the amount of money. Usually, there are three kinds of roles in the real money trade[6].

- Seller: Sell virtual property to general players to get the cash.
- Earner: Collect the virtual property by repeating specific actions in the virtual world.
- Collector: Convey the virtual property from the sellers to earners.

There are many games issuing the regulations, such as membership agreement, to ban the real money trade. However, in order to get the items and the currency in the game, the cases that players do illegal activities, such as using the game bot, are increasing. These illegal activities not only break up the balance of the game world, but also bring some trouble to the generous players. And cybercrime, such as theft of players' accounts, has also increased. For this reason, the support functions for game players are decreased, and the new players to the game are also decreased. As a result, it leads to the reduction of the revenue of the game company.

V. K-MEANS ALGORITHM

K-means algorithm is a typical kind of distance-based clustering algorithm. The algorithm uses the similarity, the object's distance, as the evaluation index. K-means algorithm regards the average value of all samples in each subset as the representative point of the cluster. In order to achieve an optimal evaluation of the clustering performance, k-means algorithm will do an iterative process that samples are divided into different categories.

The objects are determined by the distance of nearby clusters in *k*-means. It aims to obtain a high Cohesion and high independence cluster.

A. The process of k-means algorithm

Figure 1 shows the flowchart of k-means algorithm, and the steps are as follows. First, assuming that there are n objects in

total, and selecting k objects as the initial cluster centers. K is the value of clusters' number and should be given ahead. Other objects are assigned to the most similar cluster. Next, calculating the center of each cluster. Then, each object is relocated to the closest cluster. This process should be repeated until Reference value of the evaluation function converges. Usually square deviation is used as an index function. K-means algorithm's process is as follows[11]:

- Initialize the center vector $c_1, c_2, \ldots, c_k$.
- Assign samples to the clusters. Samples are assigned to the nearest cluster.
- Calculate the center vector of each cluster using the samples of the cluster.
- Repeat the step 2 and 3 until the position of the cluster center converges.

Figure 2 shows the overview of the k-means algorithm. we regard players as objects, and input the objects. Because k should be given ahead and it is usually unknown, we have done several experiments to get a better k for the study. As the algorithm shows, we need to initialize k center vectors. And then assign objects to the nearest center vector. After that, we need to calculate new center vectors again. And we should repeat the steps until the position of the cluster center converges.

Figure 3 shows the simple example of k-means algorithm. In addition, we have set k = 2. First, we need to select two center points randomly from numbers 1 , 2 , 4 , 5. Each sample is to be divided into the nearest cluster. Therefore, number 1 is divided into cluster 1, and numbers 2 , 4 , 5 are divided into cluster 2. Next, we need to calculate the new center points of the new clusters. The center point of cluster 1 is not changed, and the new center point of cluster 2 (2 , 4 , 5) is 3.67. And then, numbers 1 and 2 are divided into cluster 1(center point 1), numbers 4 and 5 are to be divided into cluster 2(center point 3.67). Next, the center point of the cluster 1 (consisting of numbers 1 and 2) is 1.5, the center point of the cluster 2 (consisting of numbers 4 and 5) is 4.5. The algorithm is to be ended, because the position of the cluster center converges.

B. Euclidean distance

In k-means algorithm , the parameter k and the function for the computing of the distance are very important factors. According to the dataset, it is necessary to select a suitable similarity function to calculate the distance between objects. In general, Manhattan distance or Euclidean distance is often used. Particularly, Euclidean distance is used more frequently. And we also selected Euclidean distance to calculate the distance between objects in this paper. Assume d is the function of Euclidean distance, then the formula of Euclidean distance of $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $x_j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ is shown as follows.

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \qquad (1)$$

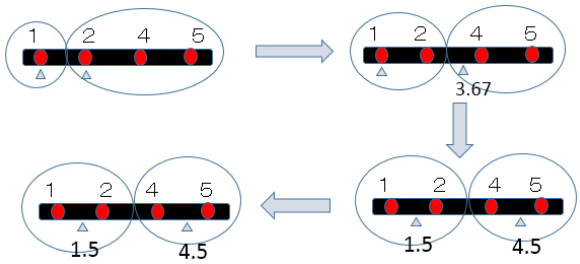


Fig. 3. Simple example of k-means algorithm.

The similarity between the samples is shown by the distance $d(x_i, x_j)$. The degree of similarity between two samples is large if $d(x_i, x_j)$ is small, and the degree of similarity between two samples is small if $d(x_i, x_j)$ is large.

C. parameter k

In k-means algorithm , the parameter k is assumed to be given ahead. However, it is very difficult to set the optimal value of k. It is usually difficult to know the optimal value of k of a dataset in many cases. And this is a faultiness of k-means algorithm. We took some values of k to do the experiments, and compared the results to get a better k in this paper.

D. Evaluation of the clustering performance of k-means

The function of squared error is ordinarily used to evaluate the clustering performance in k-means algorithm. If we are given the dataset which contains k clusters, we can represent $X = \{x_1, x_2, \ldots, x_k\}$, and $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ in each cluster. In addition, we use m_i to represent the center of each cluster, so we can get the centers $\{m_1, m_2, \ldots, m_k\}$ of the clusters $\{x_1, x_2, \ldots, x_k\}$. we show the function of square error and criterion function as follows.

$$E = \sum_{i=1}^{k} \sum_{p \in X_i} \left\| p - m_i \right\|^2 \qquad (2)$$

In formula (2), E means the sum of the squared error of the all objects in the dataset, and p represents the samples of each cluster. We use formula (2) to evaluate the performance of k-means algorithm in order to obtain high Cohesion and high independence clusters.

E. The design of the feature vector for WoWAH

In order to apply k-means algorithm to the dataset of WoWAH, we need to design the feature vector for the algorithm. According to the dataset, we regarded the Avatar ID (player ID) as the key, so different IDs are regarded as the different objects(players). And if we assume the total number of players is N, the dataset can be expressed as follows.

$$X = \{ x_m \mid m = 1, 2, \cdots, N \} \qquad (3)$$

In the formula, x_m is a feature vector that stands for each player, and $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$. In x_i, each x_{ij} represents the game playing time that is required for the level j of the avatar. For example, if x_i = (10, 25, ... ,100), x_{i1} = 10 represents that the player took 10 units of the game playing time at level 1.

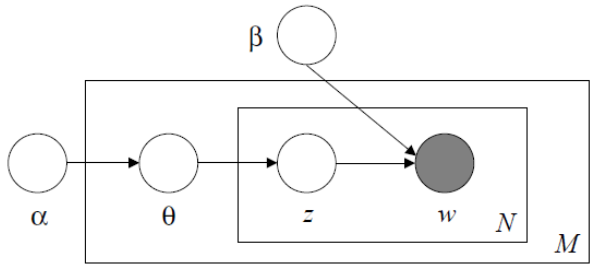


Fig. 4. Graph model of LDA.

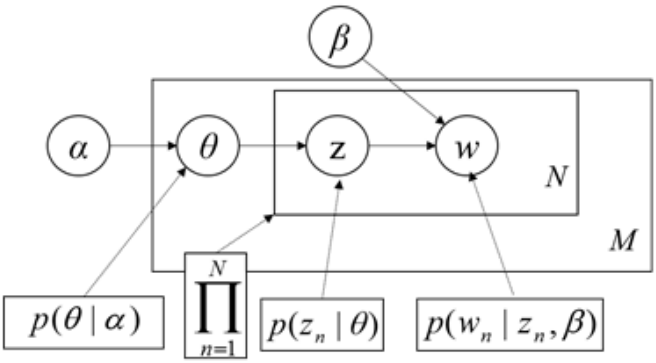


Fig. 5. Graph model of LDA with probability.

We used the function of Euclidean distance to calculate the similarity between players in this paper, and proved it could be used in k-means algorithm. We can find a player who is different from the generals by analyzing the data sample, but it will take a large amount of human labor. We will use game playing time of the players at each level to design the feature vector for the k-means algorithm. And then we can do a classification of the players, and find the irregular players whose actions were different from the generals by comparing the characteristics of the categorized players. Frauds always come with some characteristics, which can be used as the degree of similarity of the objects. Therefore, we can reduce the human labor by using the k-means algorithm.

VI. TOPIC MODEL

There are mainly two kinds of topic models, that is, Probabilistic Latent Semantic Analysis(PLSA)[12] and Latent Dirichlet Allocation (LDA). In Latent Semantic Analysis, a low dimensional space will be used to find the vector to represent the document. And in LDA, the document will be mapped to the topic space, hence we can obtain the relationship between the document and topic. In this paper, we would like to find a relationship between game playing time and levels, so we give priority to the use of LDA as the model.

If we assume that the things, which are to be searched, are consist of various subjects, we can obtain a rough topic by using the "zoom in" and "zoom out" through the article. Therefore, articles can be found by the relation among the topics. In order to find an article, you need to search for the related topics firstly, and then you can find the articles that are related to the topics. Latent Dirichlet Allocation (LDA) is a simple kind of topic model algorithm, which is based on a generative probabilistic model of a number of topics.

A. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a kind of generative probabilistic models which can transform articles into a topic model. The algorithm takes a basic idea that articles are regarded as a series of potential topics without any order. In addition, each topic is regarded as a distribution over the words. More specifically, according to the following process ,we can generate an article. Latent Dirichlet Allocation algorithm's process is as follows:

1) choose $N \sim Poisson(\xi)$. N represents the length of the article.

2) choose $\theta \sim Dirichlet(\alpha)$. θ represents the probability of each topic , α is the parameter of the Dirichlet distribution.

3) For each of the N words w_n*:*

a) Choose a topic $z_n \sim Multinomial\ (\theta)$.

b) Choose a word w_n *from* $p(w_n \mid z_n, \beta)$. $p(w_n \mid z_n, \beta)$ is a multinomial probability conditioned on the topic z_n.

In the $p(w_n \mid z_n, \beta)$, the word probabilities are parameterized by a $k \times v$ matrix β where $\beta_{ij} = p(w_j = 1 \mid z_n = 1)$. It means that β represents the probability of the words under the condition of the topic z_n. Finally, we can generate a series of worlds with β distribution under the condition of the topic z_i.

Figure 4 shows the graph model of LDA [13]. The model shows that we need to select the topic vector θ to determine the probability of each topic at first. And next, in order to generate words, we should select a topic according to the topic vector θ. We should select a word under the probability distribution of the selected topics. If we assume $p(\theta, z, w \mid \alpha; \beta)$ to be $p(w)$, we can get a set of N words w by the formula as follows.

$$p(w) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \qquad (3)$$

Figure 5 shows graph model of LDA with a probability. In order to transform articles into a topic model, we need to calculate the two parameters(α, β). We can calculate the topic vector θ of the model according to $p(\theta \mid \alpha)$. And then, we can generate the topics z_n through θ.

B. Design of the vector for WoWAH

LDA is a kind of algorithm to generate an article under the condition of topics. In order to design a vector of the topic model for the dataset of WoWAH, it needs to calculate the probability of the words under the condition of topics. That is, we have to calculate the parameter β.

We regarded an avatar ID (one player) as an article in this paper. In addition, the relation between game playing time and avatar's level was regarded as a topic in the article. In this way, we can also design the vector of the topic model as formula(3). Probability vector of topics x_m represents each player in the game, and $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$. In this sample, each x_{ij} represents the probability of generation of each word under the topic in the article(the player). For example, if $x_i = (0.02, 0.03, \ldots 0.33)$,

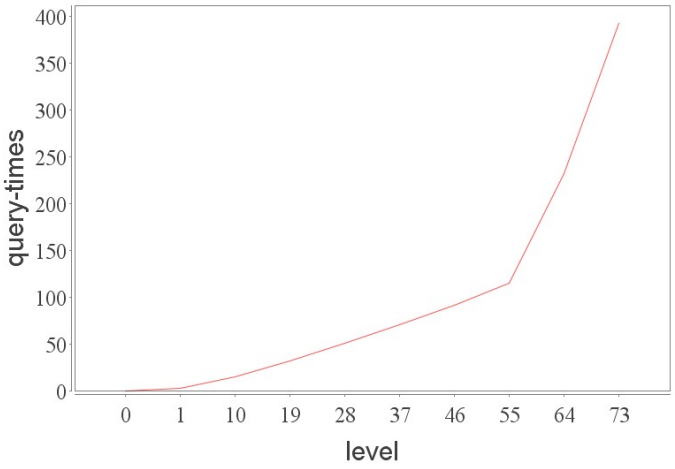


Fig. 6. Average game playing time with the level.

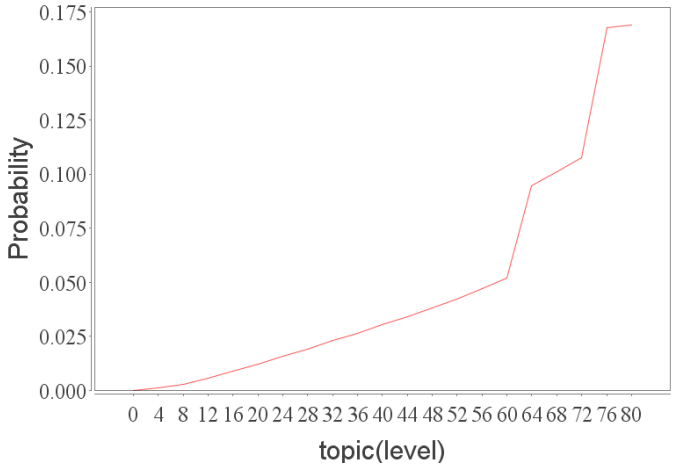


Fig. 7. Average probability under the topic(LDA).

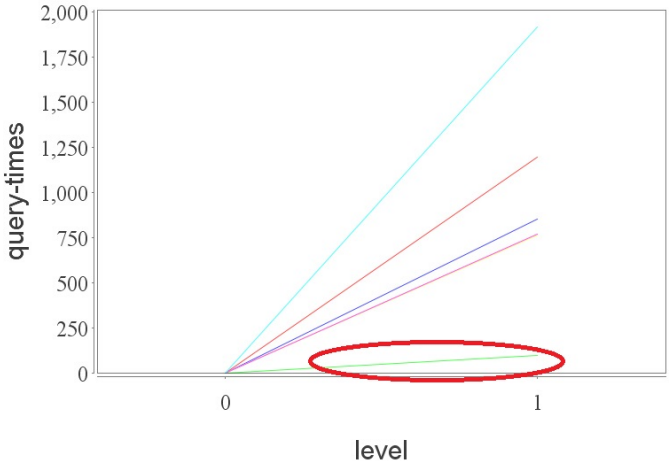
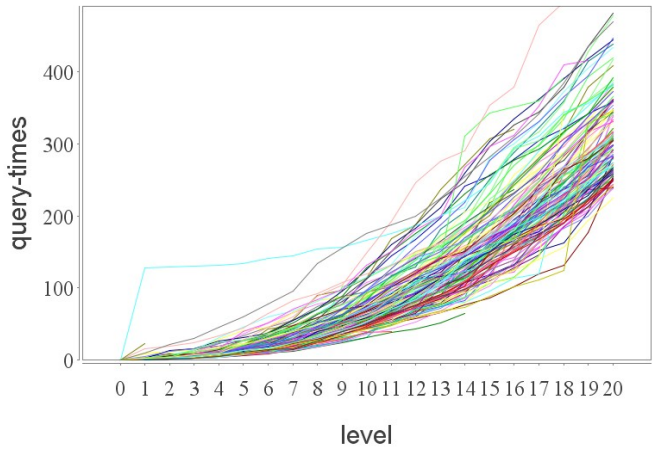
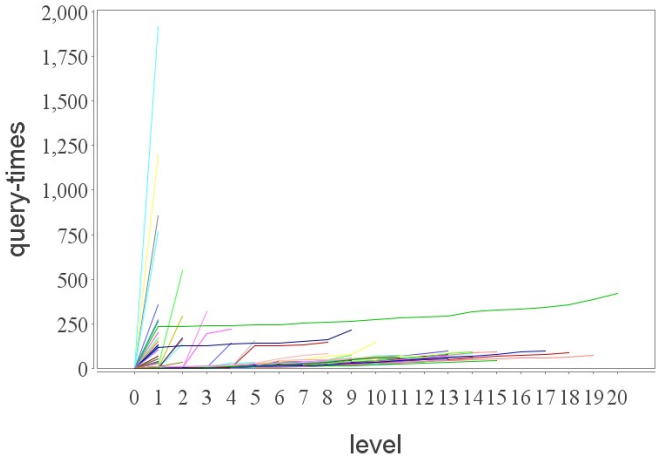


Fig. 8. Cluster D

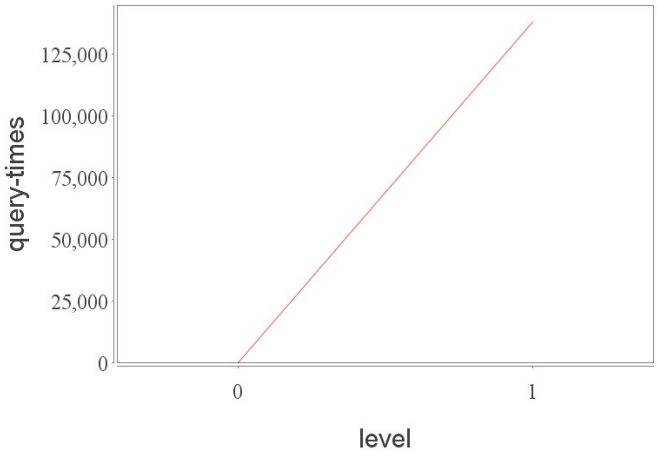
xi1 = 0.02 represents the probability of generation of a word at the condition of the topic(Level 1).



(a) Cluster A



(b) Cluster B

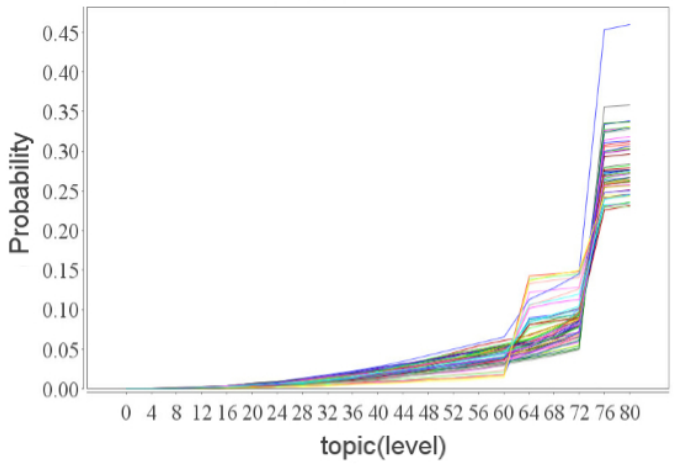


(c) Cluster C

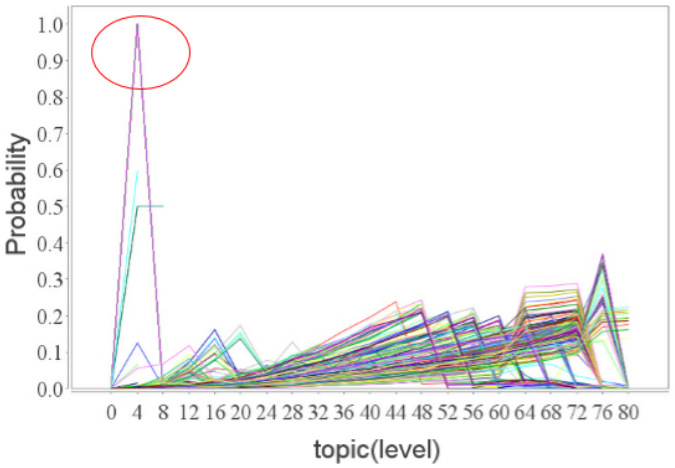
Fig. 9. Clusters obtained by k-means algorithm.

VII. EXPERIMENT

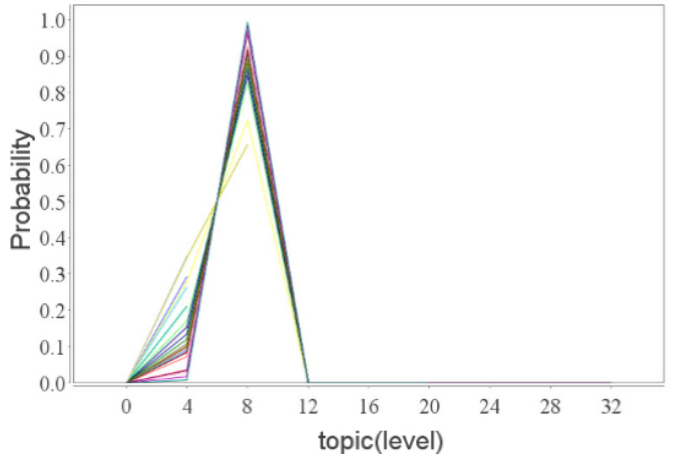
We have applied the two methods, which introduced in Section IV and Section V, to the dataset of WoWAH to detect the player group with a specific action.



(a) Cluster A



(b) Cluster B



(c) Cluster C

Fig. 10. Clusters obtained by k-means algorithm with LDA

In the dataset of WoWAH, the data from January 2006 to January 2009 is included, and there are 91,065 Avatar ID. We used avatar ID from 1 to 10000 to do the experiment, and the experiments were divided into 5 timeframes, that is, we used

TABLE III. THE STATISTICS OF EACH CLUSTERS

Avatar ID	Cluster A	Cluster B	Cluster C
1 - 2000	1095	902	3
2001-4000	1021	946	33
4001-6000	1176	799	25
6001-8000	755	1245	0
8001-10000	547	1453	0
Amount	4594	5345	61

2000 IDs for each experiment. For the most of players are normal players in the game, we can use the average of the all players as the evaluation standard to detect a player group with a specific action.

Figure 6 shows the average game playing time with the level of all players. Horizontal axis represents the levels of the avatar, and the vertical axis represents the game playing time of the players.

Figure 7 shows the average probability of generation of a word at the condition of a topic. Horizontal axis represents a topic (the levels of the avatar), and vertical axis represents the probability of the probability of the words under the condition of the topic.

In general, the level of avatars should be upgraded with the increase of game playing time, and through Figure 6 and Figure 7, we could find that the required game playing time for each level increases exponentially. Usually, the level of avatars can be upgraded in a rapid speed at the beginning, because it asks for a little empirical value. With the level upgrading, it asks for more and more experiences for a further upgrade. Therefore, the curve trends tends to increase exponentially as it shows in Figure 6 and Figure 7. Therefore, we can do a comparison with the divided clusters to find a player group with a specific action.

A. Results of using the k-means algorithm

In k-means algorithm, *k* should be given ahead. We have done several experiments to get a better *k* in this study, because the value of *k* will make a big impact on the result. In this study, the parameter k of the k-means algorithm was set from 2 to 20. This paper introduced the results when k was 4, because it has a better results than others. Although we set the k=4 in the experiment, we finally get a cluster that should be classified into the Figure 9c. As Figure 8 shows, the curve of the cluster is similar to Figure 9c. Therefore, we considered that the cluster shown in Figure 8 could be summarized in Figure 9c. But it generated an error as shown in Figure 8, which should be in the cluster of Figure 9b. The player in a red circle in Figure 8 should be classified in the cluster of Figure 9b. At last, it obtained almost as the clusters that shown in Figure 9a, 9b, 9c when k = 4.

B. Results of using the LDA algorithm

One player was considered as one article in this study. And then, we calculated the vectors of topic model by using the game playing time of players with LDA algorithm. Next, we divided players into several groups with k-means algorithm.

We introduced the results when k was 4, because it came with a better result than others. And it almost obtained the clusters that shown in Figure 10a, 10b, 10c when k = 4. The player in a red circle in Figure 10b should be classified in the cluster of Figure 10c. For the curve of marked player is similar to Figure 10c, we classified it into the cluster which shown in Figure 10c.

C. *Summary of the results*

Comparing to the average of all players, the players who shown in Figure 9a can be determined as a normal player group because the required game playing time for each level of the players in the group also increases exponentially. As shown in Figure 9b, after taking a certain level, there is no record of the game playing time of the players, so we can conjecture the players may give up playing game halfway. And as shown in Figure 9c, the amount of game playing time is also very large while the level of the players gets no change almost. The players in this group might perform an illegal action in the game. For example, the players might use game bots to perform the same task for earning gold and items.

Figure 10 shows clusters using LDA [13] before clustering. Using a topic model, we expect to find more specific clusters in the cluster in Figure 9a. Although the cluster shown in Figure 10a represents normal players, The cluster shown in Figure 10a has a special feature. The players in the cluster spend long time at a level then, they do not take time after such level. It means those players do something at some level and they could be normal players. The cluster shown in Figure 10b is similar to the cluster shown in Figure 9b, which means those players may give up the game at the early stage. K-means clustering with LDA shows better result than only k-means clustering. But, it also has errors such as a player in a red circle in Figure 10b should be classified in the cluster of Figure 10c. And by LDA, we can find the players of Figure 10c take long time at a level.

We made a statistics of the 5 times experiments. As shown in table III, there are 61 players in the cluster C, in which the players might perform illegal action in the game. In this experiment, 53.45% of players were classified in cluster C, in which players may give up playing game halfway. Moreover, the statistics of [1] has shown that 50% of players would subscribe for longer than 500 days.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we used the k-means algorithm to classify the player in order to detect the player group with a specific action in the dataset of WOWAH. Specifically, we designed a feature vector to represent a player's actions. Then we used k-means algorithm to do a classification of players based on Euclidean distance. We first did an experiment by using k-means algorithm, then we added a topic model to the experiment and obtained the results. In order to get a better parameter *k* of k-means, we have done several experiment with *k* was set from 2 to 20. At last, we got a best result when *k* was set as 4, and the result was shown in the paper.

Although there are some mistakes in the result, they might be caused by the missing samples of the dataset or the special case. Frauds usually come with some characteristics that are different from the generals. Our schemes can be extended to be more efficient if we utilize more aspects of players' activities.

We try to find a new dataset to verify the generality of the schemes in future work. And we will classify the players with a specific action based on other classification algorithms to discuss the advantages and disadvantages of the clustering methods.

REFERENCES

[1] Yeng-Ting Lee, Kuan-Ta Chen, Yun-Maw Cheng, and Chin-Laung Lei, "World of Warcraft Avatar History Dataset," In Proc. of ACM Multimedia Systems 2011, Feb 2011.

[2] Kuan-Ta Chen, Hsing-Kuo Kenneth Pao, and Hong-Chung Chang. "Game Bot Identification based on Manifold Learning," In Proc. of the 7th ACM SIGCOMM Workshop on Network and System Support for Games. ACM, 2008.

[3] Kuan-Ta Chen, and Li-Wen Hong. "User identification based on game-play activity patterns."

[4] Kuan-Ta Chen, et al. "Identifying MMORPG bots: A traffic analysis approach." EURASIP Journal on Advances in Signal Processing 2009 (2009): 3.

[5] Thawonmas, Ruck, et al. "Clustering of online game users based on their trails using self-organizing map." Entertainment Computing-ICEC 2006. Springer Berlin Heidelberg, 2006. 366-369.

[6] Atsushi Fujita, Hiroshi Itsuki, and Hitoshi Matsubara. "Detecting Real Money Traders in MMORPG by Using Trading Network," AIIDE, 2011.

[7] Itsuki, Hiroshi, et al. "Exploiting MMORPG log data toward efficient RMT player detection." Proceedings of the 7th International Conference on Advances in Computer Entertainment Technology. ACM, 2010.

[8] Mishima, Yutaro, Kensuke Fukuda, and Hiroshi Esaki. "An analysis of players and bots behaviors in MMORPG." Advanced Information Networking and Applications (AINA), 2013 IEEE 27th International Conference on. IEEE, 2013.

[9] Thurau, Christian, and Christian Bauckhage. "Analyzing the Evolution of Social Groups in World of Warcraft®." Computational Intelligence and Games (CIG), 2010 IEEE Symposium on. IEEE, 2010.

[10] Mezhvinsky, Dimitry. Looking to Sell: Assessing the Real World Value of Virtual Property. Diss. Miami University, 2009.

[11] Wagstaff, Kiri, et al. "Constrained k-means clustering with background knowledge." ICML. Vol. 1. 2001.

[12] Hofmann, Thomas. "Probabilistic latent semantic indexing." Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999.

[13] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation," the Journal of machine Learning research 3 (2003): 993-1022.