I know what the BOTs did yesterday: full action sequence analysis using Naïve Bayesian algorithm

Jina Lee
Distributed Data Processing Team
NC Soft, Inc.
Seongnam, Korea
jina.lee320@gmail.com

Jiyoun Lim
SW Future Research Team
ETRI
Daejon, Korea
kusses@gmail.com

Wonjun Cho
Distributed Data Processing Team
NC Soft, Inc.
Seongnam, Korea
wonjun.cho@gmail.com

Huy Kang Kim
Graduate School of Information Security
Korea University
Seoul, Korea
cenda@korea.ac.kr

***Abstract*— A game BOT is a major threat in the online game industry. There have been many efforts to distinguish game BOT users from normal users. Several studies have proposed BOT detection models based on the analysis of users' in-game action sequence data. These studies indicated that the analysis of users' in-game actions is effective to detect BOTs. However, they do not use sufficiently large data sets to train and test their algorithms. In this paper, we have proposed a BOT detection model that uses users' in-game action sequence data obtained with the aid of big data analysis environments. We did empirical analysis of the dataset of "Blade and Soul", the third largest MMORPG in Korea. The result shows that a large amount of sequence data leads to high accuracy.**

Keywords—BOT detection; sequence data; Naïve Bayesian classifier; online game security

I. INTRODUCTION

A game BOT is a well-crafted AI program that can play a game without any human interaction or control. Game BOTs can do harm and shorten the life cycle of the game [1]. We collaborated with NC Soft, Inc., one of the largest MMORPG service companies. We fully analyze long-term user activity logs with the aid of a big data analysis platform.

II. RELATED WORK

The studies for detecting game BOTs have been classified into three categories: server-side, network-side, and client-side [1, 2]. Network-side BOT detection assumes that traffic information, represents disparate characteristics of human players and BOTs [3]. Client-side detection methods acquire action data directly through user involvement. Server-side BOT detection methodology is based on data mining techniques that analyze log data from game servers. In general, it is hard to modify server-side logs with deliberate intent, unless hackers penetrate and compromise the game servers. Thus, server-side detection methods are more secure than client-side methods. Previous studies on server-side BOT detection methods selected features from a vast array of user actions, and applied data mining techniques [3, 4, 5, 6]. Most of these studies used one-dimensional log data due to computational complexity in spite of the fact that game BOT traces are found in multi-dimensional log data.

We choose the users' in-game action log as a dataset. Whenever a user does an action (e.g. hunting) in the game world, then the game server generates an action log for each user's action. We build the action sequence data from the dataset. The sequence data with time-stamp information is higher-dimensional. In this paper, we employ big data technology by Hadoop system for minimizing the loss of information. Additionally, Naïve Bayesian is adopted in this study because the log data used is relatively large in terms of both its count and noise levels.

III. EMPIRICAL STUDY

A. Data Description

In-game action stored in the game log represents the behavior of characters in the game. 110 kinds of log events were used for creating sequence patterns. Table 1 shows the information, including size and periods, of the training and test data. Before applying the sequence data for training, data preprocessing is needed to eliminate the insignificant data. The optimal sequence length "k" and the significant data frequency "s" are determined by the Apriori algorithm that makes explicit the relationship between data, based on the data frequency [7].

TABLE I. DATA SET USED FOR THIS STUDY

	Period (week)	# of characters	# of records	# of BOTs
Training data	2012 32nd	41,467	750 million	1,187
Test data	2012 33rd	40,364	570 million	829

The data are separated into "all sequence patterns" and "limited number of sequence patterns" to compare the performance of prediction rate based on the amount of information. Limited number of sequence patterns is decided according to the research by Ahmad et al. (2009) [3]. And twelve kinds of sequence patterns were selected.

B. Analysis based on a simple scoring algorithm

In this paper, the simple scoring algorithm used for classification [8] was applied to BOT detection by analyzing sequence data. The scoring algorithm for BOT detection is composed of three stages as shown is Fig. 1.

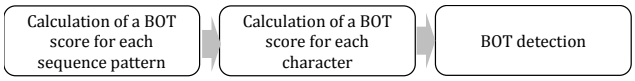

Fig. 1. Scoring Process

The BOT score for each sequence pattern is calculated by comparing normal user's counts and BOT's counts. These values are represented as a Boolean variable and score 1 means the sequence pattern occurs more frequently in BOT.

A character's BOT score is estimated by taking the average value of each character's sequence patterns. Cutoff value is decided as a minimum score of the BOT characters excepting outliers. In this case, characters having a BOT score greater than 0.76 are identified as BOT characters.

C. Analysis based on the Naïve Bayesian algorithm

The Naïve Bayesian algorithm was applied to email spam filters in the study by Androutsopoulos, et al. [9]. We applied the algorithm in our analysis as shown in (4) and (5).

$$P_i = \frac{frequency\ of\ pattern\ i\ of\ bot\ actors}{\sum_i frequency\ of\ pattern\ i\ of\ bot\ actors} \quad (4)$$

$$\mathrm{P}(B|x) = \frac{P(B)P(x|B)}{P(x)},\ P(N|x) = \frac{P(N)P(x|N)}{P(x)} \quad (5)$$

where, x: $actor$, B: $BOT\ class$, N: $Normal\ class$

IV. RESULT OF ANALYSIS

As a result, the optimal value of the length of the sequence pattern (k) and the minimum support rate (s) were decided, as shown in Table 2.

TABLE II. THE BOT DETECTION RESULT OF EACH ALGORITHM

Algorithm	k, s	The # of sequence patterns	Precision	Recall	F-measure
Simple scoring	9, 400	12	3.8%	87.2%	7.4%
		73,362	100%	78.5%	87.9%
Naïve Bayesian	4, 500	12	32.5%	68.6%	44.1%
		17,252	100%	100%	100%

139 characters, investigated by three GM (Game Master) experts, were used for the evaluation. 79 characters were proved to be BOTs with no false positives. According to the result, the amount of input data has to be large enough to include a variety of information for the training algorithm.

V. CONCLUSION

A BOT detection model based on analyzing sequence data by Naïve Bayesian algorithm is proposed in this study. We compared the results of all sequence patterns that had a value higher than that of limited number sequences. As a result, our analysis, used all kinds of sequence patterns, was more effective in identifying BOTs as the result showed the values for recall and precision were equivalent to 100% of the values that are applicable in the actual service area. Notwithstanding the contributions of this paper, in future work, the results can be compared to the results obtained by using other classifiers.

ACKNOWLEDGMENT

We would like to thank NC Soft, Inc. for providing data. This research was supported by the MSIP (Ministry of Science, ICT & Future Planning), Korea, under the C-ITRC (Convergence Information Technology Research Center) support program (NIPA-2013-H0301-13-3007), supervised by the NIPA (National IT Industry Promotion Agency). This work was also supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) and Research Development Program 2013.

REFERENCES

[1] A. Kang, J. Woo, J. Park, and H. Kim, "Online game bot detection based on party-play log analysis," Computers and Mathematics with Applications, vol. 65, pp. 1384-1395, 2013.

[2] R. Thawonmas, Y. Kashifuji, and K. Chen, "Detection of MMORPG bots based on behavior analysis," ACE '08 Proceedings of the 2008 International Conf. on Advances in Computer Entertainment Technology, pp. 91-94, 2008.

[3] M.A. Ahmad, B. Keegan, J. Srivastava, and D. Williams, and N. Contractor, "Mining for gold farmers: automatic detection of deviant players in MMOGs," CSE '09. International Conference on Computational Science and Engineering, vol.4, pp. 340-345, August 2009.

[4] A. Kang, H. Kim, and J. Woo, "Chatting pattern based game BOT detection: do they talk like us?," vol. 6, no. 11, pp. 2866-2879, 2012.

[5] J. Woo, H. Choi, and H. Kim, "An automatic and proactive identity theft detection model in MMORPGs," Appl. Math. Inform. Sci., vol. 6, no. 1S, pp. 291S-302S, 2012.

[6] C. Platzer (2011). *Sequence-based Bot detection in massive multiplayer online games* [Online]. Available: http://www.iseclab.org/papers/ICICS2011.pdf .

[7] R. Agrawal, and R. Srikant, "Fast algorithms for mining association rules in large databases," Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, Santiago, Chile, vol. 1215, pp. 487-499, September 1994.

[8] P.L. Chen, C.C. Lee, C.Y. Li, C.M. Chang, H.C. Lee, N.Y. Lee, C.J. Wu, H.I. Shih, H.J. Tang, and W.C. Ko., "A simple scoring algorithm predicting vascular infections in adults with nontyphoid Salmonella bacteremia," Clin. Infct. Dis., vol. 55, no. 2 pp. 194-200, July 2012.

[9] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, G. Paliouras, and C.D. Spyropoulos, "An evaluation of naive bayesian anti-spam filtering," arXiv preprint cs/0006013, 2000