ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (427-431), Month: April - June 2015, Available at: www.researchpublish.com

Classification of Human and Chat Bot for Server Side Security

[1]Miss. Snehal S. Deshmukh, [2]Dr. M. S. Ali

[2]M.E. (CSE) II Year, [2]Principal
[1,2] Prof. Ram Meghe College of Engineering & Management Amravati, India

***Abstract*: In this paper we examine Internet chat systems to classify the human and bot. Although chat as an application does not want huge amounts of traffic, chat systems are known to be habit-forming. This implies that catering to such users can be a promising way of attracting them, especially in low bandwidth environments such as wireless networks. A computer program designed to simulate conversation with human users, especially over the Internet chat server. chat bots often do conversations like they're a game of tennis: talk, reply, talk, reply .The flooding of chat services by automated programs, known as chat bots, poses a serious threat to Internet users. Chat bots target popular chat networks to distribute virus, spam and malware .In this paper, first conduct a study of on a huge commercial chat network. Our review capture a total of 16 different types of chat bots ranging from simple to advanced chat bots. Moreover, we observe that human behavior is very different than bot behavior. Based on the study, we see that a classification system to accurately distinguish chat bots from human users. The current available classification system consists of two components: 1) an entropy-based classifier; and 2) a Bayesian-based classifier. The two classifiers depends on each other in chat bot detection.**

***Keywords:* Bots, classification, Internet chat.**

I. INTRODUCTION

Internet chat is a popular application that enables real-time text-based communication system. Millions of people around the world use Internet chat system to exchange messages and discuss a long range of topics online. Internet chat is also a unique networked application because of its human-to-human interaction will be done and low bandwidth consumption. However, the large user base and open nature of Internet chat make it an ideal target for malicious activities [1].

The abuse of chat services by automated programs, known as *chat bots*, which poses a serious threat to online users. Chat bots have been detected on a number of chat systems, including large commercial chat networks, such as AOL [2], Yahoo! [3]–[5], and MSN [6], and open chat networks, such as IRC and Jabber. There are also reports of bots in some nonchat systems with chat features, including online games, such as *World of Warcraft* [7].

Chat systems, and especially Web-chats, are used by a sizable number of people around the world for exchanging ideas and discussions. Considering the spectrum of Internet applications, chat systems occupy the extreme, short lived interactive end, while stable, non-interactive applications like the web occupy another. Given that chat as an application does not generate a lot of traffic it is not surprising that chat systems do not contribute a large fraction of traffic to the internet. Therefore it may ask why is chatting a worthwhile target for traffic characterization. Our primary motivation is that to understand different forms of chat bots .chat offers computer mediated communication. This means that the measured"behavior" is expected to mainly depend on human behavior. Furthermore chat is used by a large number of users. It has the potential of being habit-forming [8].

Since chat systems need small bandwidth, they are an interesting and attractive wireless application. A wireless application that has gained extreme popularity in Asia and Europe is SMS (short message service) and is used by some users in a manner similar to chat. In today's Internet there are various chat systems in use which differ in a number of

aspects. Internet Relay Chat (IRC) [6] is one of the oldest systems still in use. From a high level perspective an IRC-system consists of a network of different servers which form a spanning tree among them, that is used as the backbone of the network. A client using an IRC-system connects himself to one of these servers and the messages will travel along the backbone to each of the connected servers. The IRC protocol which regulates message exchange is well defined and easy to understand. Accordingly we chose IRC as a starting point for our analysis.

II. BACKGROUND AND RELATED WORK

A. Chat Systems:

A chat **system** that allows users to communicate in real time using easily accessible web interfaces. It is a type of internet online **chat** distinguished by its simplicity and accessibility to users who do not wish to take the time to install and learn to use specialized **chat** software's. Internet chat is a real-time communication tool that allows online users to communicate via text in virtual spaces, called chat rooms or channels. There are a number of protocols that support chat [10], including IRC, Jabber/XMPP, MSN/WLM (Microsoft), OSCAR (AOL), and YCHT/YMSG (Yahoo!). The users connect to a chat server through chat clients that support a particular chat protocol, and they may browse and join many chat rooms featuring a variety of topics. The chat server relays chat messages to and from different online users. A chat service with a large user database might employ multiple chat servers. Although IRC has been existing for a long time, it has not gained mainstream popularity for this. This is mainly because its console-based interface and command-line-based operation are not user-friendly. The recent chat systems improve user experience by using graphical interfaces, as well as adding attractive features such as avatars, emoticons and audio-video communication capabilities.

B. Chat Bots:

A computer program designed for simulating the conversation with human users, especially over the internet. "chat bots often treat conversations like they are a game of tennis: talk, reply, talk, reply"

The term bot, short for robot, refers to automated programs that do not require a human operator. A chat bot is a program that interacts with a chat service to automate the human task, e.g., creating chat logs. The first-generation chat bots were designed to help operate chat rooms or to entertain chat users, e.g., quiz or quote bots. However, with the commercialization of the Internet, the main enterprise of chat bots is now sending chat spam. Chat bots deliver spam URLs via either links in chat messages or user profile links.

- Typically, a chat bot will communicate with a real person, but applications are being developed in which two chat bots can communicate with each other.
- Chat bots are used in applications such as ecommerce customer service, call centers and Internet gaming.

C. Related Work:

Dewes et al. [8] conducted a systematic measurement study of IRC and Web chat traffic, revealing several statistical properties of chat traffic. 1) Chat sessions tend to last for a long time, and a significant number of IRC sessions last much bigger than Web chat sessions. 2) Chat session inter arrival time follows an exponential distribution, while the distribution of message interarrival time is not exponential. 3) In terms of message size, all chat sessions are dominated by a large number of small packets. 4) Over an entire session, typically a user receives about 10 times as much data as he is sending.

McIntire J.P[9] several potential interrogation strategies for users and chat room administrators who may need to actively distinguish between a human and a chat bot, quickly and reliably, during distributed communication sessions. The identification problem faced by interrogators in a Turing Test, and the proposed methods and strategies might find application to and inspiration from this topic as well.

Dickerson, J.P.[10] A collection of network-, linguistic-, and application-oriented variables that could be make as possible features, and identify specific features that distinguish well between humans and bots. By analyzing a large dataset relating to the 2014 Indian election, it show that a number of sentiment related factors are of key based identification of bots, significantly increasing the Area under the ROC Curve (AUROC).

SHEEBA JAYA PRIYA[11] The classification of human, bot, and cyborg accounts on Twitter. The entropy based component uses tweeting interval as a measure of behavior complexity, and detects the periodic and regular timing that is

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (427-431), Month: April - June 2015, Available at: www.researchpublish.com

an indicator of automation. The decision maker is based on Random Forest, and it uses the combination of the features which generated by the above three components to categorize an unknown user as human, bot, or cyborg.

III. TYPES OF BOT

1) Periodic Bots: A periodic bot display messages mainly at regular interval of time. The delay periods of periodic bots, especially those bots that use long delays, it may vary by several seconds. The variation of delay period may be attributed to either transmission delay which caused by network traffic congestion or chat server delay,or message emission delay incurred by system overloading on the bot hosting machine. The posting of periodic messages is a simple but effective mechanism for distributing messages, so it is not surprising that a substantial portion of chat bots use periodic timers[1].

2) Random Bots: A random bot posts messages at random time intervals. The random bots in our data used different random distributions, some discrete and others continuous, to generate intermessage delays. The use of random timers makes random bots appear more human-like than periodic bots. In statistical terms, however, random bots exhibit quite different intermessage delay distributions than humans[1].

3) Responder Bots: A responder bot sends messages based on the content of messages in the chat room. For example ,a message ending with a question mark may trigger a responder bot to send a vague response with a URL. The vague response, in the context, may trick human users into believing that the responder is a human and further clicking the link. Moreover, the message triggering mechanism makes responder bots look more like humans in terms of timing statistics than periodic or random bots[1].

4) Replay Bots: A replay bot not only sends its own messages, but also repeats messages from other users to appear more like a human user. In our experience, replayed phrases are related to the same topic, but do not appear in the same chat room as the original ones. Therefore, replayed phrases are either taken from other chat rooms on the same topic or saved previously in a database and replayed.

5) Replay-Responder Bots: A natural next step toward statistically human-like bots is to integrate replay and responder bots. A replay-responder bot would respond to user messages based on keyword triggers, like current responder bots, and would also randomly replay human messages, like current replay bots, resulting in human-like intermessage delay and message size statistics.

6) Advanced Responder Bots: The developer of the first-generation responder bot pointed us to a more advanced, next-generation version with a highly detailed configuration, which we refer to as the advanced responder bot. The advanced responder bot is designed to be more human-like by using a large set of keywords and responses. The advanced responder bot has a much larger number of keyword triggers than earlier bots, with each keyword trigger being handcrafted.

IV. CLASSIFICATION SYSTEM

This section describes the design of existing chat bot classification system. The two main components of that classification system are the entropy classifier and the Bayesian classifier.

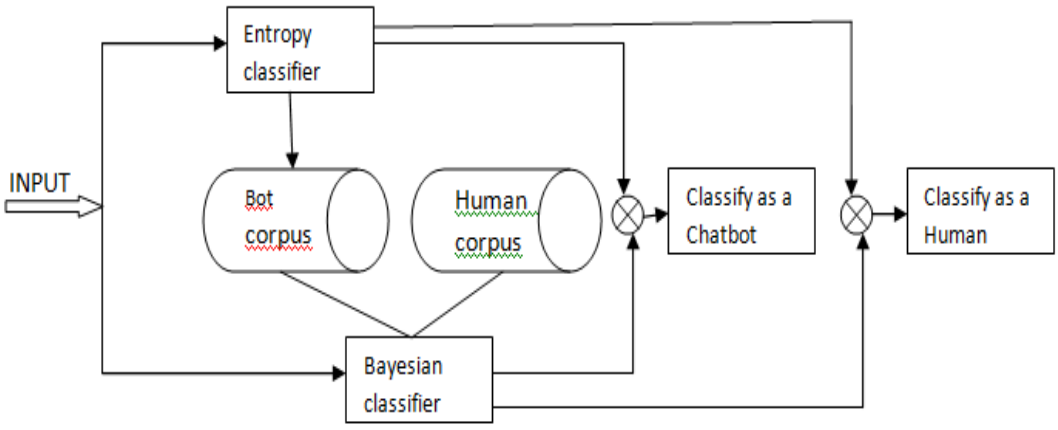

Fig 1.Classification system diagram

A. Entropy Classifier:

The entropy classifier makes classification decisions based on entropy. If either the entropy or entropy rate is low, it indicates the regular or predictable behavior of a likely chat bot. If both the entropy and entropy rate are high for these characteristics, it indicates the irregular or unpredictable behavior of a possible human. The entropy classifier measures the complexity of chat flows and then classifies them as bots or humans.

The entropy classifier helps the Bayesian classifier. The machine learning classifier requires less messages for detection and, thus, is faster, but cannot detect most unknown bots which are complex[1].

B. Bayesian Classifier:

The Bayesian classifier uses the content of chat messages to identify chat bots. Since chat messages (including emoticons) are text, the identification of chat bots can be perfectly fitted into the domain of Bayesian text classification.

A Bayesian classifier depend on the characteristics of message time and size,the entropy classifier calculates the complexity of chat flows and then classifies them as bots or humans. In contrast, the Bayesian classifier is mainly based on message content for detection. The two classifiers complement each other in chat bot detection.While the entropy classifier requires more messages for detection and, thus, is slower, it is more accurate to detect unknown chat bots. Moreover, the entropy classifier helps train the Bayesian classifier. The machine learning classifier requires less messages for detection and, thus, is faster, but cannot detect most unknown bots. By combining the entropy classifier and the Bayesian classifier, the proposed classification system is highly effective to capture chat bots in terms of accuracy and speed[1].

V. CONCLUSION

This paper first represents a large-scale study on Internet chat system. We collected information about different chat systems. we studied a total of 16 different types of chat bots and grouped them into six categories: periodic bots, random bots, responder bots, replay bots, replay-responder bots, and advanced responder bots. we found that chat bots behave very differently from human users. Although responder bots and replay bots employ advanced techniques to behave more human-like in some aspects, they still lack the overall sophistication of humans. we studied a chat bot classification system, which utilizes entropy-based and Bayesian-based classifiers to accurately detect chat bots. The entropy-based classifier exploits the low entropy characteristic of chat bots in either intermessage delay or message size, while the Bayesian-based classifier leverages the message content difference between humans and chat bots. The entropy-based classifier is able to detect unknown bots, including human-like bots such as responder and replay bots. However, it takes a relatively long time for detection, i.e., a large number of messages are required. Compared to the entropy-based classifier, the Bayesian-based classifier is much faster, i.e., a small number of messages are required. In addition to bot detection, a major task of the entropy-based classifier is to build and maintain the bot corpus. With the help of bot corpus, the Bayesian-based classifier is trained and, consequently, is able to detect chat bots quickly and accurately.. There are a number of possible areas for future work. In particular, practical deployment would raise several questions. Although aging methods used for spam filtering such as microgrooming [12] and exponential aging [13] are applicable for this study, further research is needed to determine the best approach. We also plan to investigate more advanced chat bots. For example, multiple bots could collude to forge real conversations or could perform relay attacks[14] to exploit vulnerable human users. We believe that continued work in this area will reveal other important characteristics of bots and automated programs, which could be useful in malware detection and prevention.

REFERENCES

[1] Steven Gianvecchio, Mengjun Xie, Zhenyu Wu, and Haining Wang,IEEE/ACM Transactions “Humans and Bots in Internet Chat: Measurement, Analysis, and Automated Classification“, volume -19,2011 .

[2] J. Hu, “AOL spam petitions cut both ways,” CNET News Dec. 22, 2010 [Online]. Available: http://news.cnet. com/AOL-spam-petitionscut-both-ways/2100-1024_3-1015385.html

[3] Krebs, “Yahoo! Messenger network overrun by bots,” Washington Post Dec. 18, 2007 [Online]. Available: http://blog.washingtonpost.com/securityfix/2007/08/yahoo_messenger_network_overru.html

[4] A. Mohta, "Yahoo Chat: CAPTCHA check to remove bots," Technospot.Net Dec. 18, 2007 [Online]. Available: http://www. technospot.net/blogs/yahoo-chat-captcha-check-to-remove-bots/

[5] A. Mohta, "Bots are back in Yahoo chat rooms," Technospot.Net Dec. 18, 2007 [Online]. Available: http://www. technospot.net/blogs/botsare-back-in-yahoo-chat-room/

[6] V. Hinze-Hoare, "Should cyberspace chat rooms be closed to protect children?," Comput. Res. Repository, 2004.

[7] S. Bono, D. Caselden, G. Landau, and C. Miller, "Reducing the attack surfacein massivelymultiplayeronline role-playinggames," IEEE Security Privacy, vol. 7, no. 3, pp. 13–19, May–Jun. 2009.

[8] Dewes, A. Wichmann, and A. Feldmann, "An analysis of Internet chat systems," in Proc. ACM SIGCOMM IMC, Miami, FL, Oct. 2003, pp 51–64.

[9] McIntire, J.P, McIntire, L.K.,Havig, P.R., IEEE Conference," Methods for chatbot detection in distributed text-based communications .",2010.

[10] Dickerson, J.P., Kagan, V. ,Subrahmanian, V.S., IEEE/ACM International Conference ," Using sentiment to detect bots on Twitter: Are humans more opinionated than bots?"2010.

[11] C. Sheeba Jaya Priya , K.Karthika , ISSN 2229-5518 ," Decteing online social network chat measurement,anlysis and Automated classification" ,volume -4 , March-2013 .

[12] W.Yerazunis,"Sparse binary polynomial hashing and the CRM114discriminator,"presented at the MIT Spam Conf.,Cambridge,MA,Jan. 2003.

[13] Y. Zhou, M. S. Mulekar, and P. Nerellapalli, "Adaptive spam filtering using dynamic feature space," in Proc. Int. Conf. Tools Artif. Intell. Hong Kong, China, Nov. 2005, pp. 302–309.

[14] T.Lauinger, V. Pankakoski, and E.Kirda, "Honeybot, yourman in the middle for automated social engineering," presented at the USENIX LEET, Apr. 2010