Digital Investigation 11 (2014) 349–362

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

A social graph based text mining framework for chat log investigation

Tarique Anwar [a], Muhammad Abulaish [b, *]

[a] Centre for Computing and Engineering Software Systems, Swinburne University of Technology, Melbourne, VIC 3122, Australia
[b] Department of Computer Science, Jamia Millia Islamia (A Central University), Jamia Nagar, New Delhi 25, India

ARTICLE INFO

Article history:
Received 15 February 2014
Received in revised form 3 September 2014
Accepted 13 October 2014
Available online 1 November 2014

Keywords:
Text mining
Chat logs mining
Digital forensics
Social graph generation
Cyber crime investigation

ABSTRACT

This paper presents a unified social graph based text mining framework to identify digital evidences from chat logs data. It considers both users' conversation and interaction data in group-chats to discover overlapping users' interests and their social ties. The proposed framework applies n-gram technique in association with a self-customized hyperlink-induced topic search (HITS) algorithm to identify *key-terms* representing users' interests, *key-users*, and *key-sessions*. We propose a *social graph* generation technique to model users' interactions, where ties (edges) between a pair of users (nodes) are established only if they participate in at least one common group-chat session, and weights are assigned to the ties based on the degree of overlap in users' interests and interactions. Finally, we present three possible cyber-crime investigation scenarios and a user-group identification method for each of them. We present our experimental results on a data set comprising 1100 chat logs of 11,143 chat sessions continued over a period of 29 months from January 2010 to May 2012. Experimental results suggest that the proposed framework is able to identify key-terms, key-users, key-sessions, and user-groups from chat logs data, all of which are crucial for cyber-crime investigation. Though the chat logs are recovered from a single computer, it is very likely that the logs are collected from multiple computers in real scenario. In this case, logs collected from multiple computers can be combined together to generate more enriched social graph. However, our experiments show that the objectives can be achieved even with logs recovered from a single computer by using group-chats data to draw relationships between every pair of users.

Introduction

The recent advancements in Information and Communication Technologies (ICT) are leading to several new fascinating trends in personal lives. E-communication through chat servers, Instant Messaging (IM) systems, and Internet Relay Chat (IRC) is one of the rapidly growing communication types, which initially remained popular only among the teenagers. However, due to the convenient, instant, and sophisticated nature of broadcasting and exchanging information, these days people from all walks of life find e-communication useful. Instant messaging basically refers to a client-based peer-to-peer chat discussion occurring between a small number of participants wherein the chat traffic is directly transmitted to the clients without any interruption of the server, whereas in server-based chat systems, every chat message passes through dedicated servers that direct it to the respective clients (Bengel et al., 2004). Millions of users frequently interact with their friends, family members, colleagues, or even strangers to exchange their views, thoughts, and feelings through different IM systems. Windows Live Messenger

* Corresponding author. Tel./fax: +91 11 26980014.
E-mail addresses: tanwar@swin.edu.au (T. Anwar), mAbulaish@jmi.ac.in (M. Abulaish).

http://dx.doi.org/10.1016/j.diin.2014.10.001
1742-2876/

(previously MSN Instant Messenger), AOL Instant Messenger (AIM), Yahoo Messenger, Google Talk, Skype, and Google+ Hangouts are among the popular freely available IM systems. In response to the widespread user demands, the IM systems are also popular in corporate and government organizations for news updates, notifications, marketing, and many other activities. This demand has led to the introduction of Enterprise Instant Messaging (EIM) systems that comply with security and legal aspects. IBM Lotus Sametime and Microsoft Lync Server (previously Microsoft office communications server) are the two leading EIM systems.

Most of the IM systems provide a feature to log all instant conversations and maintain history for later reference. To monitor IM chats, some commercial IM monitoring tools like WebWatcher[1] and Spector Pro[2] exist, but they provide limited functionalities. Enrichment of the functionalities of such monitoring tools to facilitate content and interaction analysis at different levels of granularity is an open research problem for the text mining community. In contrast to face-to-face communication, chat communication allows users to anynomize their identity while interacting with others. Unfortunately, this unique feature of social media has proliferated their use among anti-social and criminal persons. Hatching plots for criminal activities (like online fraudulence, hacking, drug smuggling, and terrorism), sneaking into homes to lure or cyber-bully children and opposite sexes by cyber-predators and pedophiles, and committing corporate or homeland espionage are few anti-social activities perpetrated using IM chats in a sophisticated manner by the tech-savvies (Bengel et al., 2004; Bogdanova et al., 2012; Al-Zaidy et al., 2012; Uthus and Aha, 2013). In real life, after perpetration of a crime (say a suicide which may be due to cyber-predation or drugs) at some place, the investigation team reaches the spot and investigates every piece of information that could lead to some conclusion. Due to the technological advancements of modern days, several kinds of tech-savvy gadgets are also seized from the spot, e.g., laptops, mobile phones, and memory components (microchips). Investigation of these digital gadgets opens up the research area of cyber crime investigation. This paper aims to devise a novel technique for mining huge amounts of chat logs recovered from a confiscated computer hard disk and automatically extracting critical crime-related information to assist in the investigation process.

Chat logs are usually stored as HTML files. Each HTML file represents a chat log and contains the record of a set of chat sessions along with the associated metadata. Hence, the complete chat discussion over a period of time is organized as a collection of chat sessions that are participated by the users involved in the conversation. This paper presents a complete framework to mine chat logs by applying a unified text mining approach intended to aid in cyber crime investigation. It analyzes both user interactions and conversation data together to discover their interaction patterns and overlapping interests. Although the proposed approach is generic for all kinds of chat logs, its experimental evaluation is based on chat logs archived using `Messenger Plus!` and recovered from a confiscated computer hard disk. In summary, the major contributions of this paper are as follows.

- A multi-stage chat logs pre-processing technique, including HTML tag filtering, information component extraction, noise normalization, and slang normalization, to filter out noisy and irrelevant data.
- A n-gram technique to identify candidate key-terms from message contents, and a self-customized HITS algorithm to identify feasible *key-terms*, *key-users*, and *key-sessions*.
- A *social graph* construction technique based on both chat logs metadata and message contents in which nodes represent users and links represent their ties developed through the chat interactions.
- Three possible crime investigation scenarios and a user-group identification method for each of them. *Partitive*, *hierarchical*, and *random-walk* user-group identification methods based on *k*-means, hierarchical agglomerative clustering, and Markov clustering algorithms, respectively are proposed to identify user groups with overlapping interests and user interactions patterns in different perspectives.

The rest of the paper is organized as follows. Section Challenges with informal communication presents some challenges with mining informal textual communications, followed by some related works in Section Related work. Section Problem statement states the problem addressed in this paper, and the proposed social graph based chat logs mining framework is presented in Section Proposed social graph based text mining framework. Section Experiments presents the experimental results. Finally, Section Conclusion concludes the paper.

Challenges with informal communications

In the past, chat logs have been studied for mining digital evidences, but all of them faced the common challenges posed by the noisy and informal nature of textual conversation data. The discourse of these electronic conversations is neither writing nor speech, rather it can be said as written speech or spoken writing or something unique (Kucukyilmaz et al., 2008). Their intricate sentence chunks do not follow the grammatical rules and language specific dialects that lead to the failure of language-specific parsers and traditional representation models for information extraction or distillation. Some of the major challenges that inevitably complicate the task of mining textual chat conversation data are summarized below (Agarwal et al., 2007; Kucukyilmaz et al., 2008; Aw and Lee, 2012; Schmidt & Stone):

Multilinguality: A majority of users in the world are bilingual and they frequently use multiple languages while chatting. In some cases, they use the same Latin alphabet in different languages. For example, English and Dutch both have the same set of letters and it is difficult to differentiate

[1] http://www.webwatchernow.com/Record-Instant-Messages.html.
[2] http://www.spectorsoft.com/products/SpectorPro_Windows/.