ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 5, Issue 1, pp: (105-109), Month: January - March 2017, Available at: www.researchpublish.com

Traffic Analysis of Encrypted Messaging in Various Services and Apps

Ancy Sindhya A, Maria Sheeba M

PG Student, Dept of CSE, Assistant Professor, Dept of CSE

Abstract: **The increase in usage of messaging apps enables us to collect the encrypted internet traffic. The classification of network traffic into different types of in-app service usages can help manage bandwidth and provide quality of service. Traditional approach of classification is based on packet inspection such as parsing HTTP headers. A system named CUMMA is developed for classifying service usage of messaging Apps, modelling user behavioural pattern, network traffic characteristics and temporal dependencies. The discriminative features of traffic classification can be extracted based on packet length and time delay. The clustering Hidden Markov algorithm is used for decomposing mixed-dialogs into sub-dialog which enables analyst to identify the service usages and analyse the behaviour of end user for encrypted internet traffic. CUMMA helps the mobile analyst identify the service usage and analyse end user behaviour for encrypted internet traffic, thus improving the effectiveness and efficiency of service usage classification.**

Keywords: **Encrypted Internet Traffic, In-App Analytics, Service Usage Classification, Mobile Messaging App, dialog, sub-dialog.**

I. INTRODUCTION

Mobile communication plays an important role in communication that has always been focus on exchanging of information among parties at location physically apart. Initially the mobile communication was limited between one pair of users on single channel pair. The range of mobility depends on the transmitter power, type of antenna used and the frequency of operation. With the increase in the number of users, accommodating them within the limited available frequency spectrum became a major problem. Cellular telephone systems must accommodate a large number of users over a large geographic area with limited frequency spectrum. If a single transmitter/ receiver are used with only a single base station, then sufficient amount of power may not be fixed location. The mobile data communication generally refers to the infrastructure put in place in-order to ensure that seamless and reliable communication. These would include devices such as protocols, services, bandwidth, and portals necessary to facilitate and support the stated services.

II. RELATED WORK

The increased popularity of mobile messaging Apps, such as WeChat [1] and WhatsApps [2] have become the hubs for most activities of mobile users. For example, messaging Apps help people text each another, share photos, chat, and engage in commercial activities such as paying bills, booking tickets and shopping. Therefore, service usage analytics in messaging Apps becomes critical for business, because it can help understand in-App behavior of end users, and thus enables a variety of applications. For instance, it provides in-depth insights into end users and App performances, enhances user experiences, and increases engagement, conversions and monetization. A key task of in-App usage [2] analytics is to classify Internet traffic of messaging Apps into different usage types. Traditional methods for traffic classification rely on packet inspection by analysing the TCP or UDP [8] port numbers of an IP packet or reconstructing protocol signatures in its payload. People estimate the usage types of traffic by assuming that messaging Apps consistently transmit data using the same port numbers which are visible in the TCP and UDP[12] headers. However,

there are emerging challenges for inspecting IP packet content. For example, messaging Apps are increasingly using unpredictable port numbers. Also, customers may encrypt the content of packets. In addition, governments have imposed privacy regulations which limit the ability of third parties to lawfully inspect packet contents. Moreover, many mobile apps use the Secure Sockets Layer (SSL) and its successor Transport Layer Security (TLS) as a building block for encrypted communications.

III. PROPOSED SYSTEM

To overcome the problem of encrypted internet traffic in service usage classification the CUMMA system is developed by using Hidden Markov Model. The data mining solutions is developed for classifying encrypted Internet traffic data generated by messaging Apps into different service usage types. The network traffic data of mobile messaging encode the unique patterns of both user behaviour and in-App usages. The protocol used may be UDP, which supports fast connectionless, unreliable transfer of packet. First segment the Internet traffic from traffic-flows into sessions with a number of dialogs in a hierarchical way where traffic flow denote the encrypted network traffic and session and dialog represent the segments of traffic flow in different granularity. Session is initiated when the user open the App and last until user close it. The generated internet traffic during this session is known as the dialog. Most dialogs are single type usage such as text, location sharing, voice, or stream video, while other dialogs are mixed usages.

A service usage predictor is used to classify these segmented dialogs into single-type usages or outliers. The protocol used in proposed system may be UDP, which supports fast connectionless, unreliable transfer of packet. Hidden Markov Model (HMM) is a generative model that copes with sequential data, assuming that each observation is conditioned on the hidden markov chain. HMM is used in the proposed system for classifying the service usage based on the encrypted internet traffic. Hidden Markov Chain is a Statical Markov Model in which system being modelled is assumed process with unobserved states. HMM is used to capture temporal dependencies (i.e., which produce different result over different period of time) for enhancing classification accuracy [8]. Designing a clustering Hidden Markov Model (HMM) based method helps to detect mixed dialogs from outliers and decompose mixed dialogs into sub-dialogs of single-type usage. Also CUMMA system giving power to mobile analysts to identify service usages [5] and analyse end-user in-App behaviour even for encrypted Internet traffic.

A. Traffic Segmentation:

Traffic segmentation module collects the network traffic data of different usages in messaging Apps using the data collection platform. After collecting these benchmark data, perform a two stage segmentation with these traffic-flows from coarse-grained level named session to fine-grained level named dialog. Traffic segmentation consists of two types of granularity of traffic segmentation from traffic flow to session and from session to dialog.

B. Traffic Feature Extraction:

The traffic feature extraction is used to identify their usage types and mine the discriminative features of the network traffic data from two perspectives:

- Packet Length
- Time Delay

The packet length is calculated based on the size of the packet and is measured in bytes to be transferred. Time Delay is the time taken for the packet to reach the destination [16]. Time Delay can be achieved by traffic classification by using CUMMA system. The classification for each of the probability distribution for each of the class k is given as,

$$P(k|d') = \frac{1}{B}\sum_{b=1}^{B} P_b(k|d')$$

$P_{b(k/d')}$ is the probability estimation of usage type. B denotes the number of tree. It is estimated from the ratio of the usage type.

C. Usage Type Prediction:

To predict the usage types, exploit the classification based methods to neglect the temporal dependencies thus feeding the segmented dialogs, each of which with a traffic feature vector and a reported usage type, into a robust classifier (i.e., random forest) for training and use the trained classifiers to predict the usage types of new traffic data.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 5, Issue 1, pp: (105-109), Month: January - March 2017, Available at: www.researchpublish.com

D. Outlier Detection and Handling:

To explore the temporal dependencies between consecutive user generated packets, and utilize a trained HMM model to predict the usage types of these sub-dialogs. Exploit a clustering-HMM method to segment mixed dialogs into multiple consecutive sub-dialogs of single-type usage. Since these sub-dialogs are short and only feature extraction method cannot fully describe network traffic data, to explore the temporal dependencies between consecutive user generated packets, and utilize a trained HMM model to predict the usage types of these sub-dialogs [4]. Also utilize a clustering method, then set up k centers with mean packet lengths as prior knowledge, and segment each of mixed-dialog into multiple traffic segmentations which represent a single usage sub-dialogs. Finally the trained HMM is used to predict the usage type of sub-dialogs. A system named CUMMA is used for classifying encrypted internet traffic in mobile messaging Apps by jointly modelling behaviour structure, network traffic characteristics and temporal dependencies. Encrypted internet traffic refers to the security of website traffic by encrypting the information and by using security certificates to identify and authenticate the website. It can be carried according to various parameters like port number. Each traffic can be treated differently to differentiate service provided to the user.

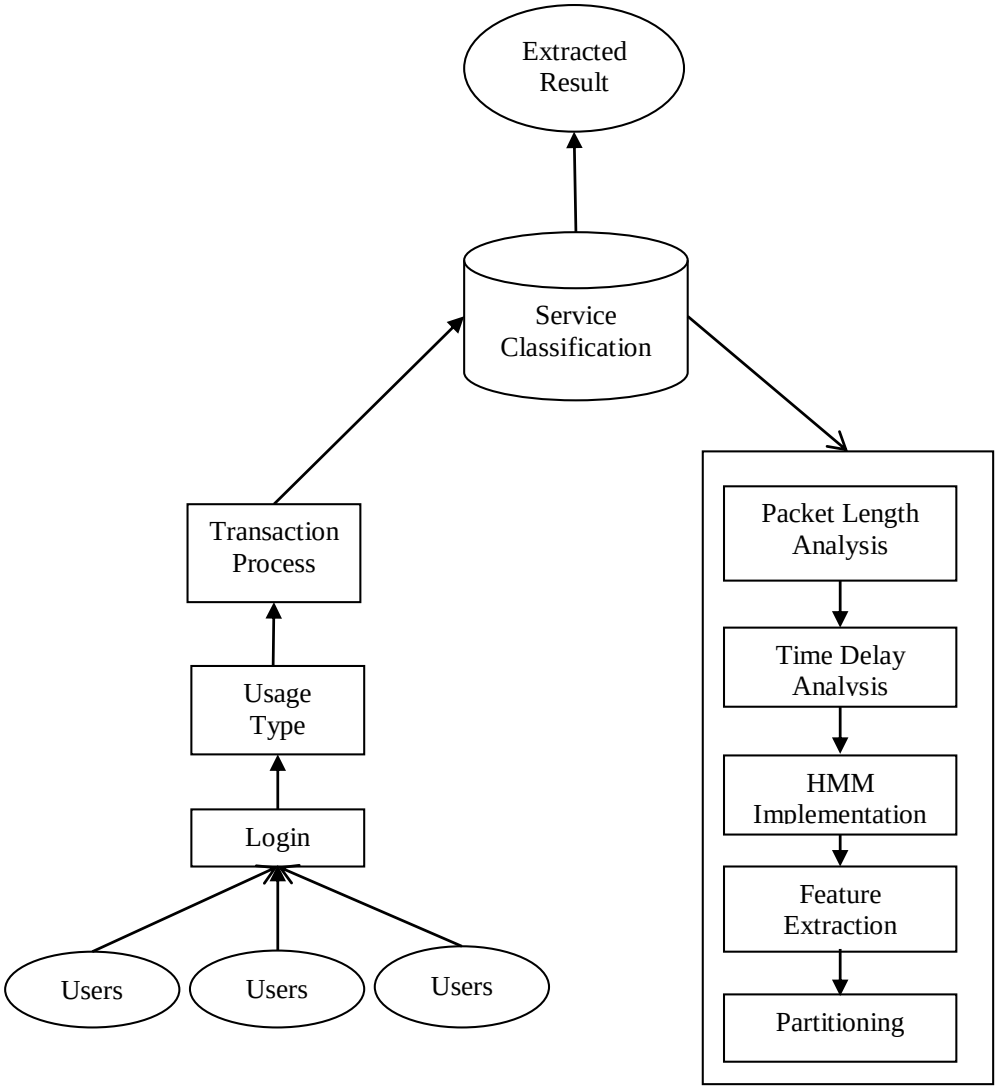

Fig.1. Architecture diagram

The Fig.1 shows the selection of usage type to identify mixed type based on encrypted internet traffic. The traffic can be analysed based on the Packet length and time delay. In this HMM is implemented such that in analysing the service usage and splitting the services into the dialog and sub dialog. First segment the Internet traffic from traffic-flows into sessions with a number of dialogs in a hierarchical way where traffic flow denote the encrypted network traffic and session and

dialog represent the segments of traffic flow in different granularity. Session is initiated when the user open the App and last until user close it. The generated internet traffic during this session is known as the dialog. Most dialogs are single type usage such as text, location sharing, voice, or stream video, while other dialogs are mixed usages. This traffic classification helps in improving the time delay based on packet size.

IV. CONCLUSION

A system is developed for classifying service usages using encrypted Internet traffic in mobile messaging Apps by jointly modelling behaviour structure, network traffic characteristics, and temporal dependencies. There are four modules in our system including traffic segmentation, traffic feature extraction, service usage prediction, and outlier detection and handling. Build a data collection platform to collect the traffic-flows of in-App usages and the corresponding usage types reported by mobile users. Then hierarchically segment this traffic from traffic-flows to sessions to dialogs where each is assumed to be of individual usage or mixed usages. First segment the Internet traffic from traffic-flows into sessions with a number of dialogs in a hierarchical way where traffic flow denote the encrypted network traffic and session and dialog represent the segments of traffic flow in different granularity. Session is initiated when the user open the App and last until user close it. The generated internet traffic during this session is known as the dialog. Most dialogs are single type usage such as text, location sharing, voice, or stream video, while other dialogs are mixed usages. Also, extract the packet length related features and the time delay related features from traffic-flows to prepare the training data. In addition, to the service usage classifiers to classify these segmented dialogs. The anomalous dialogs are detected with mixed usages and segmented these mixed dialogs into multiple sub-dialogs of single type usage. Finally, the experimental results on real world WeChat and WhatsApp traffic data demonstrate the performances of the proposed method. With this system, the valuable applications for in-App usage analytics can be enabled to score quality of experiences, profile user behavior and enhance customer care.

Advantages:

- A service usage predictor classifies these segmented dialogs into single-type usages or outliers.
- Design a clustering Hidden Markov Model (HMM) based method to detect mixed dialogs from outliers and decompose mixed dialogs into sub-dialogs of single-type usage.
- CUMMA giving power to mobile analysts to identify service usages and analyze end-user in-App behaviors even for encrypted Internet traffic.
- The extensive experiments on real-world messaging data demonstrate the effectiveness and efficiency of the proposed method for service usage classification.

Disadvantages:

- Traditional methods for traffic classification rely on packet inspection by analyzing the TCP or UDP port numbers of an IP packet.
- Messaging Apps are increasingly using predictable port numbers. Also, customers may encrypt the content of packets.
- In addition, governments have imposed privacy regulations which limit the ability of third parties to lawfully inspect packet contents.

REFERENCES

[1] Yanjie Fu, Hui Xiong, Senior Member, IEEE, Xinjiang Lu, Jin Yang, Can Chen, “Service Usage Classification with Encrypted Internet Traffic in Mobile Messaging Apps,” IEEE Transaction on Mobile Computing, 2016.

[2] Alice Este, Francesco Gringoli, and Luca Salgarelli, “Support vector machines for tcp traffic classification,” in Computer Networks, 2009.

[3] Alok Tongaonkar, Shuaifu Dai, Antonio Nucci, and Dawn Song,” Understanding mobile app usage patterns using in-app advertisements,” In Passive and Active Measurement, 2013.

[4] Eamonn J Keogh and Michael J Pazzani, "An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback," In KDD, 1998.

[5] Hossein Falaki, Ratul Mahajan, Srikanth Kandula et al(2010), "Diversity in smartphone usage," In 8th international conference on Mobile systems, applications, and services.

[6] Hyunchul Kim, Kimberly C Claffy, Marina Fomenkov et al(2008), "Internet traffic classification demystified: myths, caveats, and the best practices," In Proceedings of the 2008 ACM Context conference.

[7] Johan Himberg, Kalle Korpiaho, Heikki Mannila, Johanna Tikanmaki and Hannu TT Toivonen," Time series segmentation for context recognition in mobile devices," In ICDM, 2001.

[8] Janos Abonyi, Balazs Feil, Sandor Nemeth, and Peter Arva , "Fuzzy clustering based segmentation of time-series," In Advances in Intelligent Data Analysis V.Springer,2003.

[9] Jeffrey Erman, Martin Arlitt, and Anirban Mahanti,"Traffic classification using clustering algorithms," In Mining network data, DMIN'10, (2006).

[10] Patrick Haffner, Subhabrata Sen, Oliver Spatscheck, and DongmeiWang, "Acas: automated construction of application signatures," In Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data, 2005.

[11] Qiang Xu, Jeffrey Erman, Alexandre Gerber et al(2011), "Identifying diverse usage behaviors of smartphone apps," In Proceedings of the ACM SIGCOMM conference on Internet measurement conference.

[12] Subhabrata Sen, Oliver Spatscheck, and Dongmei Wang, "Accurate, scalable in-network identification of p2p traffic using application signatures," In Proceedings of the 13th International conference on World Wide Web, 2004.

[13] Sebastian Zander, Thuy Nguyen, and Grenville Armitage," Automated traffic classification and application identification using machine learning," In the IEEE Conference on Local Computer Networks, 2005.