Disambiguating Technique for Protecting Data in Social Network

P. Amudha

School of Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu

S.Sivakumari

School of Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu

ABSTRACT

A social network is a virtual environment powered by web technologies that enables users to publish and share all kinds of information and services with a global audience. Perhaps, social media is the vital area of the Internet, but, being open and social generates legitimate concerns about confidentiality and security. Private data is very valuable and privacy-preservation techniques provide access to sensitive contents by revealing more or less information to the users based on their credentials. This work focuses on the preservation of the output's utility and, the detection and sanitization of terms that may cause disclosure of sensitive data due to semantic correlation.

KEYWORDS-Social network, social media, privacy-preservation, sanitization

1. INTRODUCTION

A social network is made up of a set of social players and other social communications between players. Recently social networking sites such as Face book, Twitter, LinkedIn etc., have gained large popularity. Participating users of these sites form online social network, which provides sharing, organizing and finding contents and contacts. The relation between privacy and use of social network sites is very close and delicate [2]. Private information is very valuable when the information of many people gathered on social network sites. The popularity of online social network application increases serious problems about the security and privacy of their users. Privacy related with online social networking is influenced by the level of credentials of the data provided to its beneficiaries and its users. Even social networking sites that do not explicitly render the identities of their users may deliver adequate information to identify the profile of owner[1]. The commonly made mistakes that can expose an account in social network are: clicking on enticing Ads, connecting with strangers, using third party apps, exposing too much information, failing to utilize security settings, not logging out unknowingly. Growth of online social networks and publishing data in social network has led to the threat of leakage of private information of individuals. So there is need to protect owners profile and sensitive information[3].

The main aim of this paper is to prevent the risk of leakage of confidential information of individuals due to the development of online social networks and publication of social network. The rest of the paper is organized as follows: Section 2 provides literature review, section 3 presents dataset description, section 4 describes the methodology adopted, section 5 provides experimental results and discussions and finally conclusion is given in section 6.

2. BACKGROUND STUDIES

Privacy-preserving data mining has been studied widely because of the extensive dissemination of sensitive and secured information on the internet. A number of ontologybased techniques have been proposed for preserving the sensitive data. Viejo,A. et al. proposed technique to automatically detect the sensitive data in users publications and construct sanitized versions of data for the private content in social networks [9]. Sánchez, D. et al. proposed a model for document sanitization which provides a priori privacy guarantees but suffers from issues like semantic ambiguity, disambiguation and sparseness[4][5][6]. Viejo, A.andSánchez, Dproposedapproaches which tackle the threat in sensitive information of the user profiles by generating and submitting dummy queries that are related to the interests of the user.This approach uses social networks to gather more precise user profiles that allow better personalized service while offering alike, or even better, level of practical privacy[8].Velásquez, J. D,et al., proposed a model which enables the determination, within each particular context, of those actions which can threaten individual privacy[7].

3. DATASET DESCRIPTION

For experimental purpose twitter dataset has been taken for sentimental analysis. The NASA twitter dataset is obtained from the Twitter through API; it is very easy to collect millions of tweets for training. It contains around 476 million tweets. Twitter data's which contains @ symbol and #, !and URL's tags are removed during preprocessing.Fig 1 presents the twitter dataset which contain references to other tweeters (@<user>) and hash tags #<tag>, punctuations (!), quotes and Fig 2 presents the preprocessed twitter dataset.

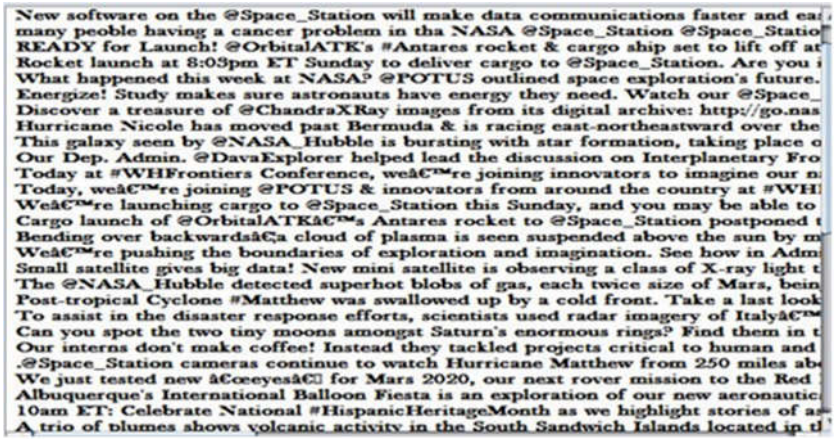
New software on the @Space_Station will make data communications faster and ea
many people having a cancer problem in tha NASA @Space_Station @Space_Statio
READY for Launch! @OrbitalATK's #Antares rocket & cargo ship set to lift off at
Rocket launch at 8:03pm ET Sunday to deliver cargo to @Space_Station. Are you
What happened this week at NASA? @POTUS outlined space exploration's future.
Energize! Study makes sure astronauts have energy they need. Watch our @Space_
Discover a treasure of @ChandraXRay images from its digital archive: http://go.nas
Hurricane Nicole has moved past Bermuda & is racing east-northeastward over the
This galaxy seen by @NASA_Hubble is bursting with star formation, taking place o
Our Dep. Admin. @DavaExplorer helped lead the discussion on Interplanetary Fro
Today at #WHFrontiers Conference, weâ€™re joining innovators to imagine our n
Today, weâ€™re joining @POTUS & innovators from around the country at #WH
Weâ€™re launching cargo to @Space_Station this Sunday, and you may be able to
Cargo launch of @OrbitalATKâ€™s Antares rocket to @Space_Station postponed t
Bending over backwardsâ€¦a cloud of plasma is seen suspended above the sun by m
Weâ€™re pushing the boundaries of exploration and imagination. See how in Adm
Small satellite gives big data! New mini satellite is observing a class of X-ray light t
The @NASA_Hubble detected superhot blobs of gas, each twice size of Mars, bein
Post-tropical Cyclone #Matthew was swallowed up by a cold front. Take a last look
To assist in the disaster response efforts, scientists used radar imagery of Italyâ€™
Can you spot the two tiny moons amongst Saturn's enormous rings? Find them in t
Our interns don't make coffee! Instead they tackled projects critical to human and
.@Space_Station cameras continue to watch Hurricane Matthew from 250 miles ab
We just tested new â€œeyesâ€ for Mars 2020, our next rover mission to the Red
Albuquerque's International Balloon Fiesta is an exploration of our new aeronautic
10am ET: Celebrate National #HispanicHeritageMonth as we highlight stories of a
A trio of plumes shows volcanic activity in the South Sandwich Islands located in t

Fig1: Sample Twitter Dataset

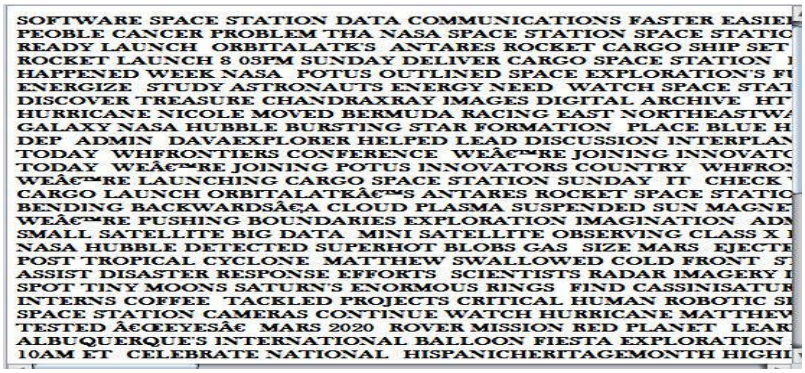
SOFTWARE SPACE STATION DATA COMMUNICATIONS FASTER EASIE
PEOBLE CANCER PROBLEM THA NASA SPACE STATION SPACE STATIC
READY LAUNCH ORBITALATK'S ANTARES ROCKET CARGO SHIP SET
ROCKET LAUNCH 8 03PM SUNDAY DELIVER CARGO SPACE STATION
HAPPENED WEEK NASA POTUS OUTLINED SPACE EXPLORATION'S FU
ENERGIZE STUDY ASTRONAUTS ENERGY NEED WATCH SPACE STAT
DISCOVER TREASURE CHANDRAXRAY IMAGES DIGITAL ARCHIVE HT
HURRICANE NICOLE MOVED BERMUDA RACING EAST NORTHEASTWA
GALAXY NASA HUBBLE BURSTING STAR FORMATION PLACE BLUE H
DEP ADMIN DAVAEXPLORER HELPED LEAD DISCUSSION INTERPLAN
TODAY WHFRONTIERS CONFERENCE WEÂ€™RE JOINING INNOVATO
TODAY WEÂ€™RE JOINING POTUS INNOVATORS COUNTRY WHFRO
WEÂ€™RE LAUNCHING CARGO SPACE STATION SUNDAY IT CHECK
CARGO LAUNCH ORBITALATKÂ€™S ANTARES ROCKET SPACE STATIC
BENDING BACKWARDSÂ€¦A CLOUD PLASMA SUSPENDED SUN MAGNE
WEÂ€™RE PUSHING BOUNDARIES EXPLORATION IMAGINATION ADV
SMALL SATELLITE BIG DATA MINI SATELLITE OBSERVING CLASS X
NASA HUBBLE DETECTED SUPERHOT BLOBS GAS SIZE MARS EJECTE
POST TROPICAL CYCLONE MATTHEW SWALLOWED COLD FRONT TA
ASSIST DISASTER RESPONSE EFFORTS SCIENTISTS RADAR IMAGERY I
SPOT TINY MOONS SATURN'S ENORMOUS RINGS FIND CASSINISATUR
INTERNS COFFEE TACKLED PROJECTS CRITICAL HUMAN ROBOTIC SI
SPACE STATION CAMERAS CONTINUE WATCH HURRICANE MATTHEW
TESTED Â€œEYESÂ€ MARS 2020 ROVER MISSION RED PLANET LEAR
ALBUQUERQUE'S INTERNATIONAL BALLOON FIESTA EXPLORATION
10AM ET CELEBRATE NATIONAL HISPANICHERITAGEMONTH HIGHL

Fig 2: Preprocessed Twitter Dataset

4. METHODOLOGY

4.1 PRIVACY-PRESERVING GENERALIZATIONS

The privacy requirements of the user are attained when setting up the system into the user's environment. In this approach, the type of interactions between users of the social network contains three different privacy levels:

1. User1(FULL ACCESS): Can see all the posts
2. User2(PARTIAL ACCESS): Can see the generalized posts
3. User3(NO ACCESS): Can see any of the posts

The linguistic analysis is done on the input message to be published by the user and extracts potentially sensitive terms with regard to the thresholds of each privacy level.Replacing a sensitive term with a generalization (e.g., cancer - >disease), we are lowering the amount of information disclosed to a certain type of readers while retaining an amount of its semantics/utility.

Assuming privacy requirements with n levels { L 0 ,…,L n-1 } and their corresponding n thresholds { T L0 ,…,T Ln-1 }, for each term t in a certain message m to be published do:

- The system will not do anything and t will be published as is, so that readers in the level L0 (or higher) can use it.
- The system acquires the most informative generalization g0(t) from the databases in use such that TL0 ≥ IC(g0(t)).
- The system acquires the most informative generalization gi(t) from the databases in use such that TLi ≥ IC(gi(t)).
- The system acquires the most informative generalization g0(t) from the databases in use such that TL0 ≥ IC(g0(t)).

The sensitive words are identified by using linguistic analysis and then by using sanitization the words are generalized and then visible to user who is permitted with the partial access and then for no access. The terms or generalizations categorized in the lowermost privacy level (*L0*) will get available in the social network.

4.2 PROTECTING SENSITIVE DATA

The sets {*S1,…,Sn-1*} contain the termsor generalizations (i.e., sanitized versions) that are accessible for the users belonging to each privacylevel. The terms in{*S1,…,Sn-1*} are encrypted that each reader will get only the *Si* linked to the privacylevel *Li.*The set of encrypted elements will be used by authorized readers to obtain the right cryptographic keys and get the corresponding protected terms. It is to avoid availability issues or problems inherent to distributed solutions;the system stores the protected information in the servers of the social network.

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

Input is syntactically analyzed and protected messages in social networks and their percentage of information preservation shows that User has published a message that most of the members are affected by cancer. Therefore, according to the permission, *L2* readers can know everything, thus *TL2* =∞. *L1* readers can only know the generalized term which, according to the generalizations of CANCER in the knowledge base's in use would correspond to "disease" *TL0* = *IC*(*"disease"*) = - log2(126E6/17E9) = 7,1.finally, any external entity, which in this case would be only the social network operator that corresponds to *L0*, will not even know the type, that is *TL0* = 0,. This does not allow any information accessed by the user. Reader L0 only know encrypted message.

Terms and generalizations published in the social network and stored in the Si corresponding to each privacy level Li; $ states that the generalization for Si is the same as such already published. The last row shows the size (in bytes) of each Si to be stored in attached images. Table 1shows, “In NASA many people having disease problem”, in this tweet sensitive word cancer is generalized to disease and it is stored in images to publish to other users.

Table 1. Size of Each Si to be stored in attached images

PUBLISHED	S1	S2
Many	$	Many
People	$	People
Having	$	Having
Disease	Disease	Cancer
Problem	$	Problem
NASA	$	NASA
Bytes to store	15 bytes	59 bytes

Ontology based disambiguating technique is adopted which gives highest score for each word and the different levels of access provided for the user is shown in Fig 3, 4 and 5. Fig 3 shows that when the user is provided with full access, they can view full tweets. Fig 4 shows that the **u**ser provided with partial access can view generalized tweets by using ontology technique wherein Fig 5 indicates that user permitted with no access can view only encrypted message.

Also a comparison of sanitization and ontology method in terms of Precision, Recall, F-Measure is done and is shown in Fig 6.Precision value of sanitization method is 80% and of Ontology method is 83%. Recall value of sanitization is 70% and ontology is 77% and an F-Measure value of ontology is 79.8% and sanitization is 74.6%.

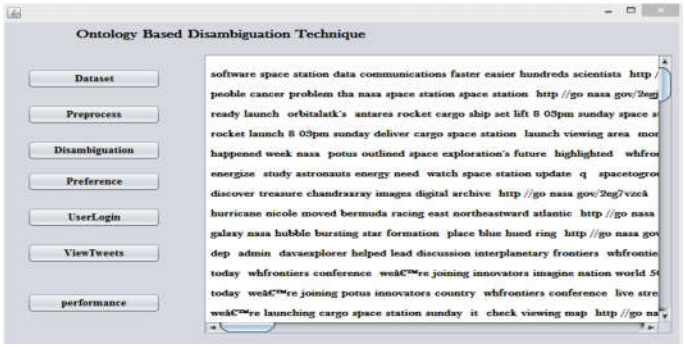

Fig 3: Disambiguating Technique for Full Access

Fig 4: Disambiguating Technique for Partiall Access

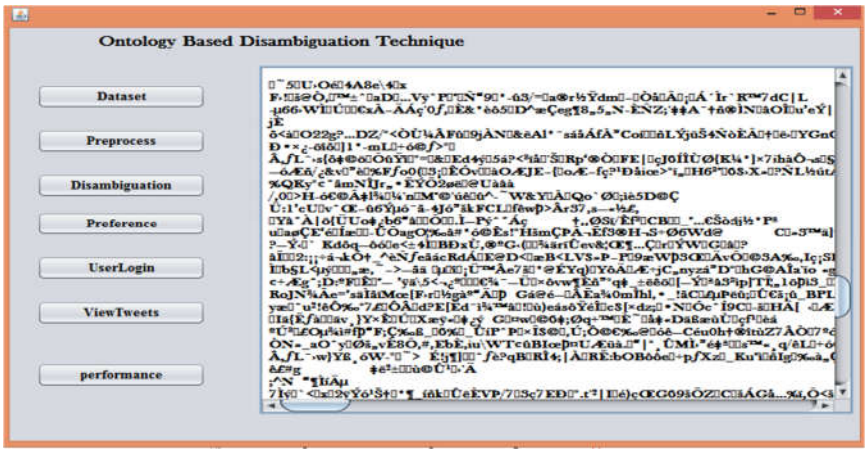

Fig 5: Disambiguating Technique for No Access

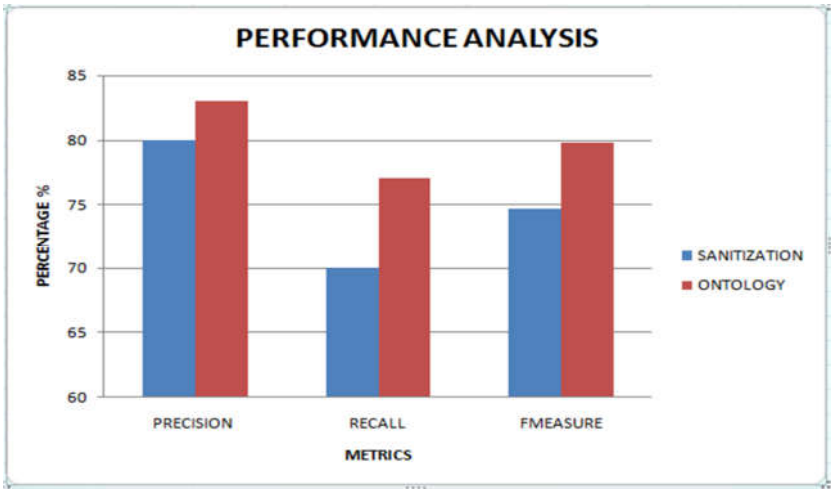

Fig 6: Overall Performance comparison

6. CONCLUSION

In the recent years, privacy of data in online social networks data has been of ultimate concern. In this proposed work, the preservation of the output's utility and, the detection and sanitization of terms are focused it may cause disclosure of sensitive data due to semantic correlation. In addition to that ontology's are used to retrieve the generalization. The experiments are conducted in terms of recall, precision and F-measure. From the experimental results it is proved that the proposed method has high precision, recall and F-measure than the existing methods.The scope of the privacy preserving in online social networks still to be explored.

REFERENCES

[1] Viejo, A., & Sánchez, D. 2014. Profiling Social Networks to Provide Useful and Privacy-Preserving Web Search. Journal of the Association for Information Science and Technology, 65, 2444-2458.

[2] Viejo, A., Sánchez, D., &Castellà-Roca, J. 2012. Preventing automatic user profiling in Web 2.0 applications. Knowledge-Based Systems, 36, 191-205,2012.

[3] Sánchez,D., Batet, M., & Viejo, A. 2014. Utility-preserving privacy protection of textual healthcare documents. Journal of Biomedical Informatics, 52, 189-198.

[4] Sánchez,D., Batet, M., & Viejo. 2014. A.Utility-preserving sanitization of semantically correlated terms in textual documents. In Proceedings of the 2nd international conference on Information Sciences, 279, .77–93.
[5] Sánchez, D.,&Batet, M. 2016. C-sanitized: A privacy model for document reduction and sanitization. Journal of the Association for Information Science and Technology, 67, 148–163.
[6] Anandan,B.,&Clifton,C. 2011. Significance of term relationships on anonymization. InIEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent AgentTechnology - Workshops, 253–256.
[7] Velásquez, J. D. 2013. Web mining and privacy concerns: Some important legal issues to be consider before applying any data and information extraction technique in web-based environments. Expert Systems with Applications, 40, 5228–5239.
[8] Batet, M. ,Harispe, S. , Ranwez, S. , Sánchez, D. , &Ranwez, V. 2014. An information theoretic approach to improve semantic similarity assessments across multiple ontologies. Information Sciences, 283, 197–210.
[9] Carminati, C. , Ferrari, E. , Heatherly, R. , Kantarcioglu, M. , &Thuraisingham, B. 2016. Semantic web-based social network access control. Computers & Security, 30, 108–115