Privacy Aware Semantics Based Prevention of Information Spread in Online Social Networks

[1]Aishwarya Kannan, [2]Bhuvaneswari Anbalagan

[1]PG Student, [2]Teaching Fellow
[1,2]Department of Computer Technology, Madras Institute of Technology,
[1,2]Anna University, Chennai, India.

***Abstract*: Online Social Networks (OSN) consists of enormous amount information shared by the OSN users. Hence information privacy has become a major concern in the OSN. In OSN, users share certain sensitive information which are easily leaked and disclosed by other users in the social network. This is because the users lack the knowledge of the access control mechanisms available in order to prevent such information leakage and are also unaware about data privacy. Therefore there is a need to automatically protect the information disclosed in the OSN. Previously a trust based analysis was done by measuring the trust between the users in the OSN with respect to their privacy awareness. But quantitatively measuring the trust for users in such a largely growing and dynamically changing social network has become a tedious task. Moreover the sensitivity of the information shared was not taken into consideration. Hence an automatic semantic annotation method was proposed which analyses the sensitivity of the information and provides access control by sanitizing the sensitive terms present in the information. The sensitivity analysis was improvised by considering the relationship strength between the users concerning the particular sensitive topic. Experiments were conducted and it was further analyzed that the proposed methodology is more effective and also the performance of the proposed methodology scales well linearly. Hence an effective and an automatic method to prevent information leakage in the OSN is proposed which correlated with the privacy settings provided by the user.**

***Index Terms*: Online Social Networks, Information Dissemination, Relationship Strength, Semantic Annotation, Sensitivity Analysis**

I. INTRODUCTION

Social network refers to a social structure that consists of individuals called nodes. These nodes are tied together by certain types of dependency such as friendship, common interest, dislike, knowledge etc. Social network analysis considers social relationships and formulates it as a network consisting of nodes and ties [8]. Nodes refer to individual actors in a network, and ties represents the relationship between the actors. The graph that occurs are usually taken to be complex. Social networks have many levels. In a more precise way, social networks is a map of certain specific ties and the nodes to which concerned individual is connected and is the individual's social contact.

The recent emergence of Online Social Networks (OSN) is similar to the traditional social networks. The unique characteristics of an OSN are in terms of formation, evolution and analysis. Online social networks have the property of forming and evolving in a bottom up manner with respect to the individuals concerned. An individual in a OSN is not location restricted, thereby enabling access to users which were previously restricted. This enhanced network reachability enables users to form social communities and then share and dissipate information to other users. OSN users have the possibility to directly communicate with each other and this is done in order to share their knowledge base [6]. The advent and popularization of online social networks made available a enormous amount of data from social interactions.

Social Networks are prone to various drawbacks. Some of the challenges on SNA are

- **Fragmentation of the social graph:** The very first drawback to the usage on social network data is fragmentation of the available data into various proprietary and the involvement of networks that are closed.
- **Discovering communities and analysis:** Social networks are vast and they continue to grow with the involvement of new users. Hence the topology of a social network keeps changing as a very fast pace.
- **Providing security by analysing social networks:** The information that are extracted from social networks are prone to attacks including insider and outsider attacks [21].
- **Social and Ethical Issues in a Social Network:** Child abuse, shared information leakage are some the issues concerned when social networks are taken into the scene.
- **Dimensioning media applications by traffic prediction:** In order to dimension the servers of the media and network resources to restrict congestion and improving QoE information such as media consumption have been utilized.
- **Spam and other adversarial interaction:** Spam and advertisements will continue growing with the increase in users and data production.

Privacy is considered to be a major concern in social networks [19, 23].
Privacy has a variety of meanings and plenty of definitions. This will vary from information privacy to personal privacy and most of it concentrates the web [31]. Privacy harms are caused by most of the OSN users and visitors [2]. The following are the different threats concerning privacy in OSN

- **Strangers can view personal information**: Though users understand OSN as protecting their data from being leaked, in the real true world it is not so [29]. Even security has many blunders that can involve leaking the information to unknown users [32].
- **Unable to hide the information shared from a particular friend or a group:** At times a person might want to hide certain information from a particular friend or from a groups of friends [18]. But this is difficult in an OSN unlike in real life where the same can be achieved easily.
- **Other users who post information about some other user :** Though a user is very selective and careful in ensuring that the right information is shared to the right users or friends , the user cannot avoid or control what other friends share or post in OSN which can even be offensive [14].

In OSN users can build their own social groups of friends and join social groups and communities. Though predefined user group settings for the users present in a social circle of friends exist, these setting sometimes prove to be inefficient. Information shared by the user are often containing sensitive information related to the user. These information can be easily leaked by individuals in a group causing privacy concerns [25]. Previous methodologies to provide privacy protecting methods for users are most often involved on a trust based measurement. In a trust based analysis, the trust level of the users are quantitatively measured by considering the privacy trust and the privacy awareness. Such methods of measuring the trust of users in a dynamically changing network topology and in a network where most of the threats are inside threats prove to be complex and inefficient.

Further, the privacy setting for users in user groups involves manually setting the privacy for each and every information being shared. But such access control are not very efficient as they are not understood completely and it is also difficult to be managed by the users [4]. These access control rights do not prove to be effective as it does not take into consideration as to whether the information contains sensitive terms or not [16]. Moreover the sensitivity of the information is also not considered in these methodologies.

Though trust has been used as a measure to determine privacy levels, precisely measuring the trust level quantitatively in a dynamic social network topology becomes a difficult and a complex task. Therefore a methodology is to be proposed which does not consider the trust level calculation and mitigates the complexity involved with trust calculation. In addition to this, in order to mitigate the difficulties caused by manually setting up access rights for each user type concerning each resource, the methodology must automatically analysis the sensitivity of the information shared and also the relationship strength between the users by analysing the semantics in other words the meaning of the information being shared [15].

The rest of the paper is organized as follows. Section II contains related work. Section III presents the proposed methodology. Section IV shows the implementation procedure. Section V contains the results and their analysis. Section VI summarizes and concludes the paper.

II. Related Work

Privacy in Online social Network has been identified as a major concern as far as the security in concerned. This is mainly because the information shared by a user in a online social network is being leaked through various forms. Users who are considered as friends of a particular use may also tend to leak the information to other users [34]. Various research works have been carried out in providing privacy to the information shared in social networks in order to protect the information from being leaked. Some of the methods include trust aware privacy protection framework, community structure for controlling information sharing, Information privacy concern with respect to peer disclosure etc. Zeng et al proposed the TAPE framework for OSN privacy. The TAPE framework concerns with Trust Aware Privacy Evaluation framework where information privacy is provided to personal information being shared in OSN. Amit et al came up with a methodology to control information sharing using community structure in online social networks which is a community centric information flow control [3]. Kaze Wong et al presented the trust and privacy exploitation in Online Social Networks. They have formulated a survey to present the drawbacks in various security mechanisms in Online Social Networks and identify the attack methods [20]. Annika gave the basis of online privacy concern highlighting the privacy concerns of different groups of different users. The study which was based on samples gives a view of how privacy concerns are taken into account in different social networks [5].

The privacy concern in OSN is pointed out as a very essential factor which is based on the application under consideration. It is based on the fact when trust between users is more the concern for personal information becomes less. In a dynamically changing network topology, calculating the trust of users becomes a difficult task because most of the threats arriving are considered to be insider threats. Hence privacy has to be provided in a way that considers the semantics of the information being shared. Malik Imran et al devised privacy control in social networks using automatic semantic annotations. Here access control is taken care by analysing the sensitive terms in the information being shared and automatically annotating the words with other terms that represent more generic words [23]. Sanitization of the words was originally proposed by David Sanchez et al. Here, The authors have proposed a novel methodology to remove sensitive words and terms present in information that was shared by looking into the information that was shared semantically [12]. This sanitization is very similar to a method known as redaction which usually avoids utility issues by generalizing all the sensitive words [10].

David Sanchez et al designed the ontology based semantic similarity here the evaluation of the semantic likeness between words is taken into consideration [11]. In the methodology semantic similarity which exploits the knowledge sources as base to perform all the estimation is done. Ontology plays a vital role here as they form the basis of semantic web. Ala Atrash et al gave the basics of notes and annotation as information resources in a social network site [1]. The model represents an original semantic model where notes and annotations are used as information resources. Social network users have the problem of not understanding the privacy settings provided for them. Based on this, Miriam Bartsch et al proposed a solution to control the Facebook. The method was an analysis of online privacy literacy [26]. In order to provide effective and responsible communication in a social network, users must decide between the one holding the personal information and the one disclosing the personal information.

Some of the other methodologies proposed for providing information privacy in online social networks can be described as follows. Jin Chen et al proposed the privacy of information concern about peer disclosure in Online Social Networks. In Online sites, a user's information can be co-owned and disclosed by peers [17]. This indicates the essentiality of information concern with respect to peer disclosure in OSNs. Privacy breach of location information also occurs in the OSN [6,7]. Nan Shen et al proposed an efficient and privacy preserving location sharing mechanism. A novel method known as BMobishare was proposed which was a security enhanced privacy preserving location sharing mechanism using boom filters [27]. Tabitha James et al proposed the dual decision model for online social networks. Privacy control in OSN must consider both the desire to control information and interaction which leads to dual privacy decision for users in OSN [33]. Francesco Buccafurri et al developed a methodology for the analysis preserving protection of user privacy against leakage of information of likes in social network [13].

Lorenz Schwittmann et al proposed the privacy preservation in Decentralized Online Social Networks. Storing personal data in large Online Social Networks (OSNs) gave rise to many types of privacy problems and untrustworthy users [22]. Therefore researches have proposed a decentralized architecture to create OSNs with certain privacy concern imposed on them. Natalia Criadoa et al proposed the implicit contextual integrity in online social networks [28]. Constantinos Patsakis et al came up with the distribution of privacy policies over multimedia content across multiple online social networks [9]. The relationship strength is to be measured for the type of information shared between the users in online social networks are been studied in detail.

III. PROPOSED METHODOLOGY

The main objective is to provide a methodology to protect the privacy of the user concerning the information shared. Information shared in the online social network are prone to various threats. These threats are mostly concerned with insider threat which consists of people in the friend list. Analysing the trust of the users is not an effective strategy to set the privacy rules for a user.

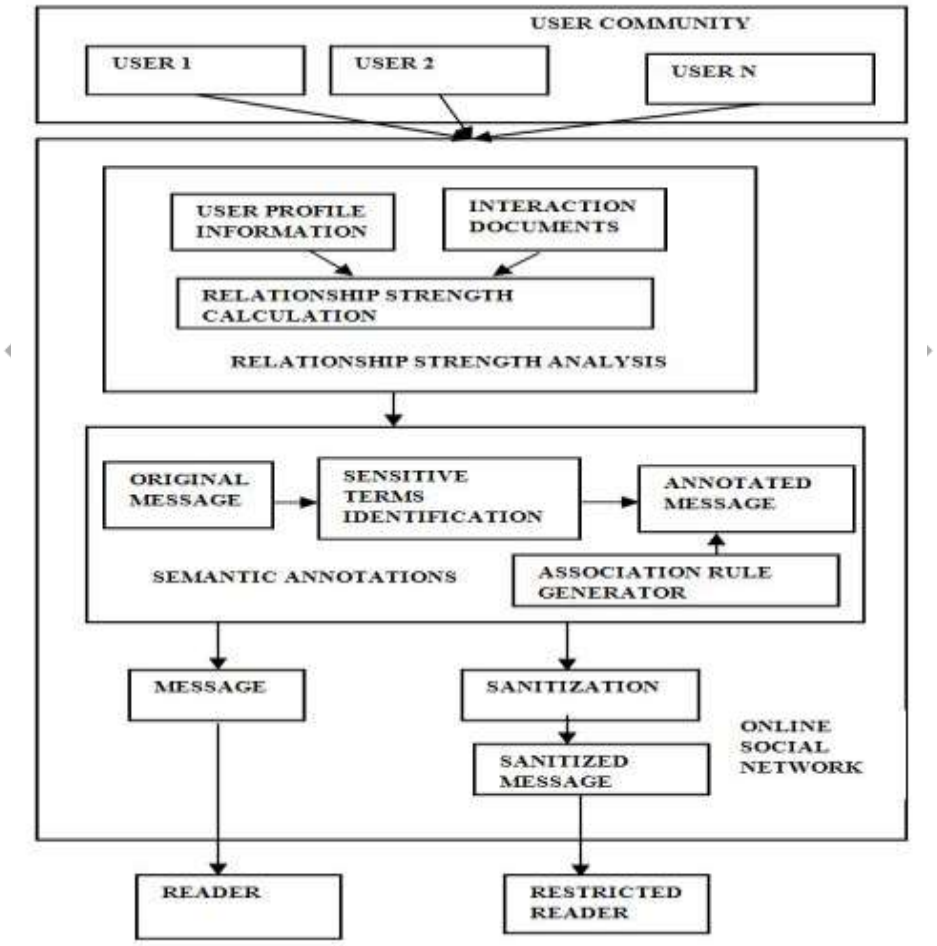

Figure 1. Overall Architecture

Hence the sensitivity of the information is to be analysed to precisely know the sensitive data in the information shared by the user. This analysis is achieved by semantically annotation the sensitive words in the information shared with related words. In addition to this, the relationship strength of the user with relation to the particular sensitive topics are also considered. The Relationship strength is calculated for individual documents shared concerned with the sensitive topic identified. Combining the predefined access rules consisting of the access levels, the sensitive terms and the User categories along with the relationship strength a more precise and an accurate privacy protection scheme related to sensitive data in the shared information can be provided. The proposed architecture shown in fig. 4.2 consists of the relationship strength calculation and the semantic annotation. This is done in order to provide an automatic and a precise determination of privacy for users in user groups mitigating the need to analyse the trust measure between the users and hence decreasing the complexity.

During relationship calculation, few sensitive topics are identified and given as input. The profile information of the user is used to measure the similarity between the two users. The interaction documents between the users are taken into consideration and a relatedness measurement is done for similar documents concerned to a particular sensitive topic. The measurements are then used to calculate the overall relationship strength between the users relating to particular sensitive topics. During semantic annotation, the information is analysed to identify the sensitive terms. The sensitive terms are then annotated with related terms representing general terms by performing semantic analysis and semantic disambiguation.

These terms are then used while sanitizing the information for the restricted users. The annotated message is then stored in a annotated message database for future reference. The access rules accordingly for every user are set while the users register with the social networking site. The users specify the terms that they consider as sensitive along with the User Categories. Therefore the sensitive topic, the User Categories, the access levels and the relation strength for the particular sensitive topic forms the access rules of users. This access rule then forms the basis for checking the user privacy setting. The access rules are then compared with that of the User category of the user accessing it. The relationship strength is also determined if it crosses a mean threshold level for the particular sensitive topic. If the access rule is acceptable for the particular user, the shared information is shared as such. If not, the message is sanitized with a more generalized term relating to the sensitive term is replaced. This sanitization of messages prevents the information from being leaked out to unauthorized users and that is being done automatically.

The proposed architecture is divided into the following components

- Sensitive Activity based Relationship Strength Analysis
- Sensitivity analysis using Semantic Annotation

Sensitive Activity based Relationship Strength Analysis

A similarity based measurement is done by using a similarity vector of information related to the sensitive activity. The data sets are then crawled to obtain the documents. The documents are identified as to belong to a user by a vector. The documents are then clustered based on Latent Dirichlet Allocation (LDA) algorithm to group all the related documents together. The clusters are then analysed to measure the relatedness of the cluster with respect to the identified sensitive activity. The same is proceeded for the individual documents. With these the strength of the interaction documents are analysed. Finally, using the profile similarity measure and the strength of the interaction activity, the relationship strength is calculated with respect to a particular sensitive activity.

Sensitivity Analysis using Semantic Annotations

The sensitive words in an information being shared are first identified. These sensitive words are mostly proper nouns which are identified by named entity recognition and the common nouns are identified by parts of speech tagging. These nouns which are identified are semantically analysed to identify the related terms. Semantic Disambiguation of the noun phrases occur to identify the noun that comes under aggregated semantically similar noun. The identified nouns are then annotated with the original message to fetch the annotated message. The annotated message is then stored in the annotated message database. The access rules are set accordingly. The access rules are compared to check if the user is authorized or not authorized and to sanitize the message accordingly. In order to facilitate the annotation process, an association rule generator is used which mines association rules based on the noun and the already annotated text. These rules are then learned by the system for prediction of annotations later on.

Factors Calculated for Determining Relationship Strength

The various factors that are computed for calculating relationship strength are as follows.

The cluster identified is related to a particular sensitive topic is given by the relatedness between the cluster and the sensitive topic.

Relatedness (c,st)=∑frequency of the word * NGD(word,st)

where c is the cluster and st the sensitive topic.

The similarity measure between the document and the cluster is given by the similarity distance measure which is defined by the cosine distance. The similarity measure can be given as

$$\text{Sim(d,c)} = \frac{WF^m . WF^n}{||WF^m||.||WF^n||}$$

where WF is the word frequency for m^{th} cluster and n^{th} document.

The interaction document between the users that were identified is then used to find out their relatedness with respect to a particular sensitive topic. The relatedness measure is then given as

Relatedness (d,st)= relatedness between cluster and sensitive topic * Sim (d,c) where d is the document and st the sensitive topic

The similarity between two user profiles is represented as a similarity vector U_i and U_j with respect to the number of attributes.

$Sim^{ij}=(s_1^{ij}, s_2^{ij}....s_n^{ij})$

The strength of the interaction document with respect to a particular sensitive topic between two users U_i and U_j is given as $St_{IA}=\sum$ (relatedness between document and sensitive topic) * u_{di} * u_{dj}

The overall relationship strength between two users with respect to a particular sensitive topic is given as

$RS_{i,j}$=Sum of profile similarity + Strength of the interaction document

Calculate Relationship Strength

The relationship strength between the users is analysed by comparing the profile information and the interaction documents based on a particular sensitive activity. The interaction documents are clustered using Latent Dirichlet Allocation (LDA) algorithm. For each of the cluster so obtained the word frequency is calculated and maintained as a vector. This word frequency and the normalized Google distance are used to calculate the relatedness between the cluster and a particular sensitive topic. The next step involves measuring the similarity of the document in the cluster using cosine similarity.

ALGORITHM : CALCULATE RELATIONSHIP STRENGTH
Input : User U_i and User U_j , Topics A
Output : Relationship Strength $RS_{i,j}$
Obtain dataset values DS={ds_1, ds_2.....ds_n}
for each ds_i € DS
get the Documents D={d_1, d_2.....d_n} from DS
end for
for each document d_i in D
Perform Clustering to generate clusters C={c_1, c_2..... c_n}
end for
for each cluster c_i in C
Record the word frequency in a vector WF_i
Calculate the relatedness of each cluster in c_i with all topics A
Assign the Topic with the highest relatedness to the c_i
end for
for each document d_i in D
Calculate the similarity of each document in d_i with all topics A
Calculate the relatedness of each document in d_i with all topics A
end for
Compute UD={$ud_{i,j}$} for U_i and U_j
for each attribute P do
Calculate Similarity Vector $Sim^{ij} = (S_1^{ij}, S_2^{ij}..... S_p^{ij})$
end for
Calculate Strength of Interaction Activity St_{IA}
Calculate the relationship strength $RS_{i,j}$ using St_{IA} and Sim^{ij}
return $RS_{i,j}$

This is followed by relatedness measurement between the document and the sensitive activity field using the relatedness of the cluster with respect to the sensitive topic and the similarity between the document and the cluster. The profile information is used to calculate the similarity vector between the users .The interaction activity strength is measured the relatedness of the cluster with respect to the sensitive activity field and the user document relation. Using the similarity vector and the interaction activity strength the relationship strength is calculated between the users for a particular sensitive topic.

Automatic Annotation of Messages

Automatic annotation of messages involves the identification of sensitive terms. The sensitive terms are usually nouns in an information and hence the proper nouns and the common nouns. The proper nouns are identified by a Named Entity Recognition and the common nouns are identified by Part of Speech tagging. Finally all similar nouns are identified and the related nouns are also formulated. The semantic similarity of all the nouns are then calculated and the most similar and generic ones are taken as the semantic annotations as a taxonomic structure. The semantic annotations are then stored in an annotated message table along with the identified sensitive nouns.

ALGORITHM : AUTOMATIC ANNOTATION OF MESSAGES
Input : Message M
Output : Annotated message $M_{annotated}$
for every word W in the text do
identify the proper nouns P
save the proper nouns identified
identify the common nouns C
save the common nouns identified
end for
for all nouns N in C U P
identify similar noun phrases N_{sim}
identify related noun phrases N_{rel}
end for
for all nouns in N_{sim} U N_{rel}
Calculate Semantic Similarity
Collect the Semantic Annotations
Store the annotations for N
end for
Append N and A as $M_{annotated}$
return $M_{annotated}$

Accessing a Message

When a reader requests to access a message the annotated message is first retrieved. The user category and the privacy requirement of the user is also fetched. The relationship strength between the reader and the owner is checked to see if it is more than the threshold value. If such a condition is satisfied, then the rule for the user is retrieved relating to the sensitive topic. The rule is checked if the access level is the same as the user's privacy requirement. If so, the message is displayed to be read for the user. If not, the message is sanitized with more general terms and a sanitized version of the message is given to the user. This ensures that the sensitive terms are hidden from unauthorized users automatically and thereby improving privacy concerns.

ALGORITHM: ACCESSING A MESSAGE
Input : Request to access a message
Output : Message m
get the Annotated message $M_{annotated}$
get the Reader classification $R_{classify}$
get the Privacy Requirement $Priv_{req}$
identify the Topic A for $M_{annotated}$
get the Relationship Strength $RS_{i,j}$ for A
if $RS_{i,j}$ > threshold t do
for all sensitive topics st_i in ST={st_1, st_2.....st_n} , uc_i in User Category UC={uc_1, uc_2.....uc_n} and al_i in Access Level AL={al_1, al_2.....al_n} do
define Access Rules $rule_i$=<st_i, uc_i, al_i>
end for
Identify the user type Ut
Identify the rule of Ut as $rule_{ut}$
Check AL in $rule_{ut}$
Compare $M_{annotated}$ with AL
if AL in $rule_{ut}$
return M
else
for all sensitive terms in M
Replace sensitive terms by generic terms
end for
Store the modified message $M_{modified}$ as M
end if return M

IV. IMPLEMENTATION DETAILS

The proposed system was implemented using Java. The implementation comprises of the relationship calculation between the social network users and the semantic analysis of the information shared by the users. This requires social network datasets which are used as inputs for the purpose of analysis. Twitter an online social networking site that has been used by various users to send and read messages which are called as tweets. The tweets are posted by the users either privately or publicly. These tweets are prone to various security threats even if shared privately since the friends themselves cannot be trusted. The datasets were collected using Twitter4J API which is a java based open source library. This API crawled up to 11,000 followers and friends of a user to denote relationship between users. About 1,000 tweets were then crawled pertaining to certain users and also the privacy setting of certain users were crawled. These datasets were then used as inputs during implementation. The tweets are the shared messages by the user. The friends and follower list provides the accessibility set by the user for another user and the privacy setting denotes the security of a user's timeline.

The relationship strength was then calculated using APIs that were used to calculate the relatedness measure such as the WS4J (Wordnet Similarity for Java) and the clustering was done using the JGibbLDA API. Taking all the words in 1000 tweets that were crawled, Clustering was then performed with 30 iterations for training. The four identified topics are climate, entertainment, health and others. The topics are later to be attached to each cluster in order to show that a particular cluster represents a particular topic. For implementation the JCN metrics were taken into consideration from the WS4J API as they gave a more accurate result than other metrics with respect to the words available in the document. These were then used to calculate the relationship strength which was used during semantic analysis.

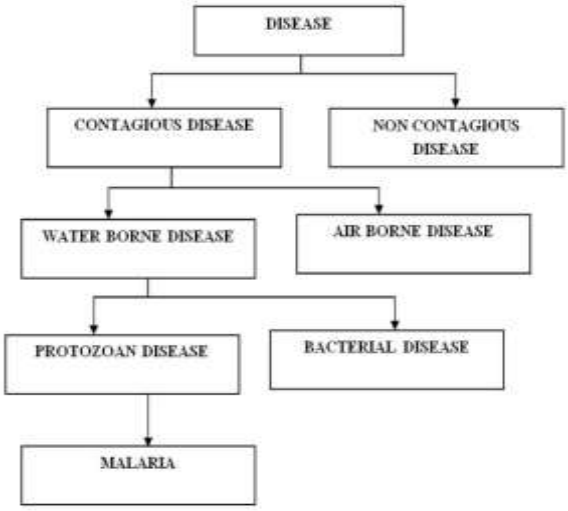

Figure. 2 Taxonomic Categories

The semantic analysis of the message shared by the user was analysed using the Stanford NLP API. A Java implementation of a named entity recognizer is the Stanford Named Entity Recognizer. The main purpose of this API is that it has the functionality of labelling a sequence of words which are present in the information shared as names of things such as a person, country name, company name etc. The Stanford CoreNLP API was used for POS tagging. Further the analysis process was done using the DBPedia knowledge base using the query tool SPARQL. The query processing was enabled using Apache Jena API which provides the query engine for SPARQL.

For the sensitive term Malaria, a query was written for while retrieving all resources with the name Malaria in them, resources such as Malaria (disease) where fetched as they contain the text Malaria. The most appropriate meaning is then disclosed by comparing the semantic distance of the word with other nouns given in the text. The related resources are then identified and its taxonomic categories are retrieved as shown in figure. 2 and annotated with the word accordingly. The annotated sensitive words are then stored in the database along with their respective annotations. Hence look up of annotations for the sensitive term becomes easier.

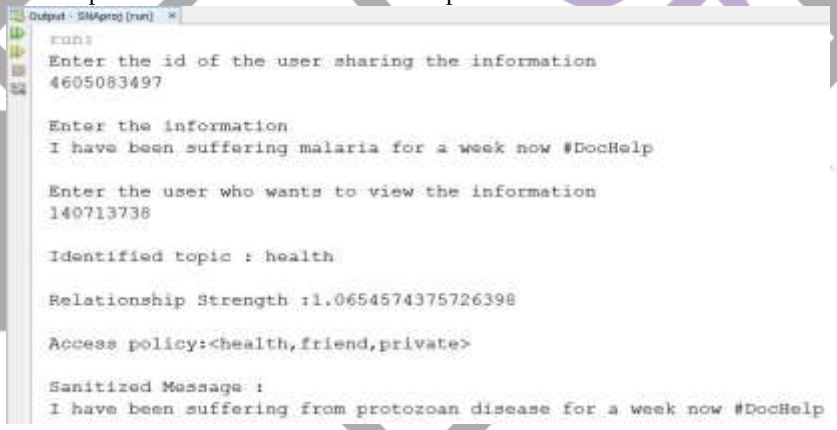

Figure.3 Sanitized Message after comparing the Relationship Strength and the Access Policies.

In figure. 3, the final message displayed is shown as the sanitized message. The relationship strength between the two users is calculated for the identified topic which is ' Health '. The strength was found to be below the threshold strength. The threshold is calculated as the mean value of the highest relatedness in each of the cluster. The below given formula denotes the threshold. Let there be N clusters. Since the threshold value is not satisfied and the access policy denotes the topic 'Health' as the sensitive topic, the message that is requested to be viewed gets sanitized replacing the sensitive term 'Malaria' with its generalized term ' protozoan disease '.

V. Experimental Results and Performance Analysis

The implementations were done and the results were analysed. Parameters were evaluated to analyse the efficiency of the proposed work in comparison to the existing works. Performance parameter, accuracy parameter and Recall parameter were taken into consideration for the estimation of the results. Performance is defined as the time taken for a particular process to take place and Accuracy is the accuracy with which the process is done. This can be explained as the number of correctly classified inputs to the total number of inputs. This is also defined as a probability distribution value of how accurate the defined classification is said to be. The Accuracy value is supposed to be high to denote a rightly classified set of inputs. Recall is defined as the percentage of users for whom the shares messages where correctly hidden to the total number of users for whom the terms are to be hidden as detected manually by a human expert. A high recall value is expected to denote that the information is rightly hidden from the intended users. The first graph shown in figure 4 was drawn for the LDA clustering to identify the right number of cluster for the input dataset.

During clustering the comparison was done with different number of clusters for the topic field assignment and its accuracy was observed. Accuracy is defined to be the number of correctly classifies words to the total number of words. The analysis was done by taking the number of clusters in the X-axis and the accuracy in the Y-axis. The accuracy value was obtained. The accuracy value along with the different number of clusters was then plotted in the graph.

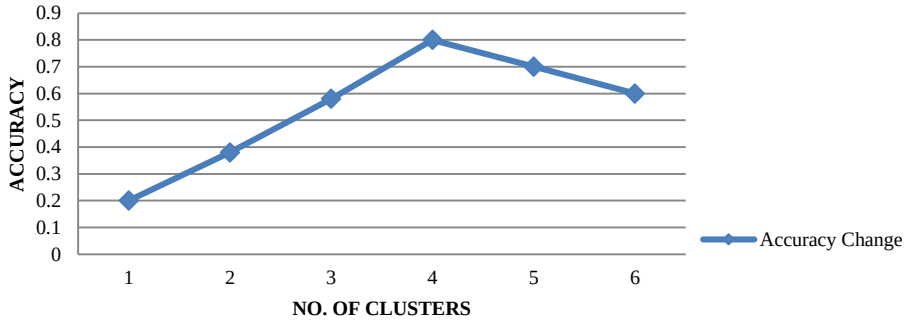

Figure. 4 No. of Clusters Vs Accuracy

The overall accuracy reaches the maximum when the number of cluster is around 4 considering all the 1000 tweets of the dataset into account. When the cluster number is less, too many documents cluster in a particular cluster and hence clustering accuracy with respect to the topics is very low. But it increases gradually and at a point it reached its maximum after which the accuracy values drop when there are too many clusters as very less documents are present in each of the cluster. From the above analysis the best number of cluster was obtained to be 5.

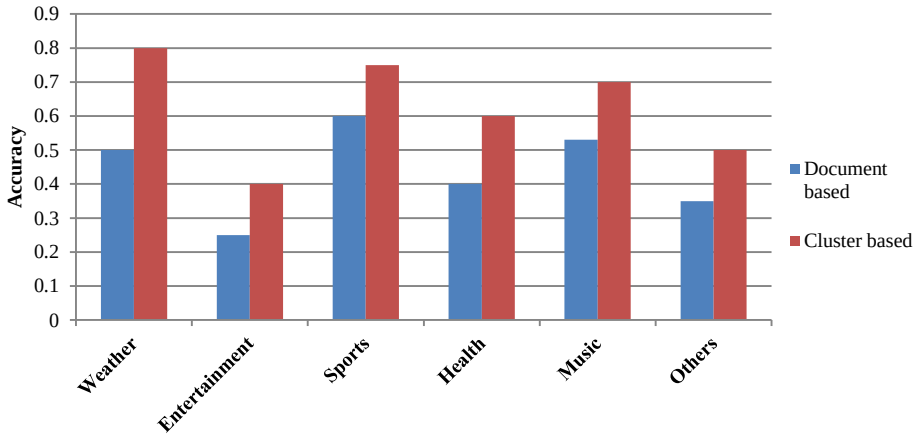

Figure 5. Comparison between Document Based Accuracy and Cluster Based Accuracy

Assigning the topics to the documents becomes a very important task. Documents can be assigned a topic directly by measuring the similarity of the document with the topic directly and assigning the topic to the documents directly and then the document is said to be belonging to that topic. The documents can also be assigned topic based on the clusters. The documents are initially clustered. The clusters are assigned a topic and then the topic is assigned to the document. To obtain an analysis on the activity field assignment for the document, the comparison was done between the cluster based activity assignment and the document based activity assignment. In order to do this, the accuracy values were taken into account. From figure 6 so obtained it was observed that the cluster based topic assignment gives a better accuracy for all the topics when compared to the document based topic assignment.

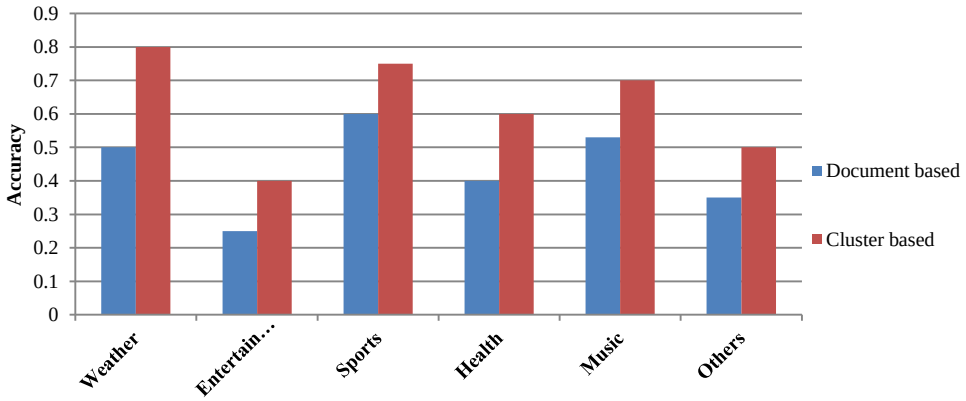

Figure 6. Comparison between Document Based Accuracy and Cluster Based Accuracy

The relationship strength was thus calculated by taking into account the relatedness of the document with the sensitive topic. The time taken for the calculation of the relationship strength is to be analysed. Hence different pairs of users were taken into consideration with different number of documents and the time taken for the relationship strength calculation was then observed and a graph was plotted.

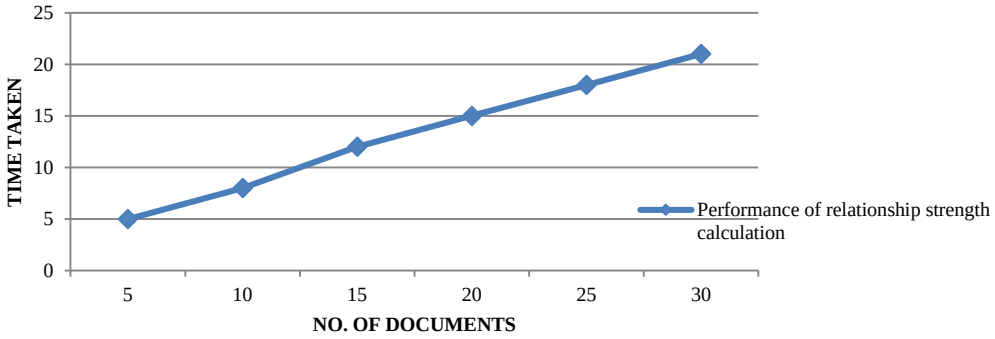

Figure. 7 No. of Documents Vs Time Taken for Relationship Strength Calculation

From figure.7 it can be observed that the time taken increases with the increase in the number of documents of the user. The graph that was plotted is shown to increase linearly with the increase in the number of documents. Thus it can be concluded that the relationship strength does not cause much overhead and can hence be calculated with a linear increase of time when the number of documents under consideration increases. Thus the bottleneck caused with the relationship strength calculation is not high and hence the system performance in an average manner. To analyse the time taken for annotation process, the time taken for the annotation process was is to be taken into consideration. A graph was plotted with the number of noun phrases taken in the X-axis and the time taken for the entire annotation process in the Y-axis. From the graph in figure 7 so obtained it was observed that the graph increases and scales linearly with the increase in the number of noun phrases. Also due to the implementation of the rule generator the time taken has slightly improved when compared to the original graph. Thus the performance of the annotation process has greatly improved and the difficulties with slow processing of the lengthy information are also reduced drastically.

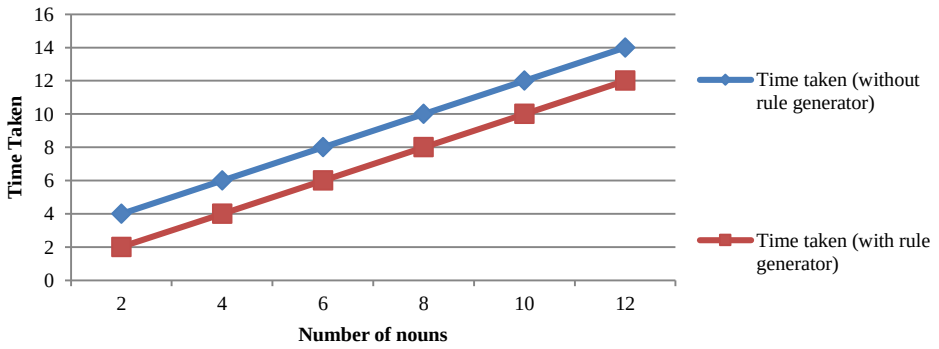

Figure 8. Comparison Between The Time Taken For Annotation With Rule Generator And Without Rule Generator

To perform a comparison of the proposed technique in detecting sensitive terms and hiding them with respect to that of the existing technique, the term recall was taken into account. Recall measures the number of users for whom successfully hidden terms was done by the system. In order to do this the percentage of correctly hidden terms that are sensitive and hidden for the number of users to the total number of users is tabulated. The improvised system also takes into consideration the relationship strength of between the users.

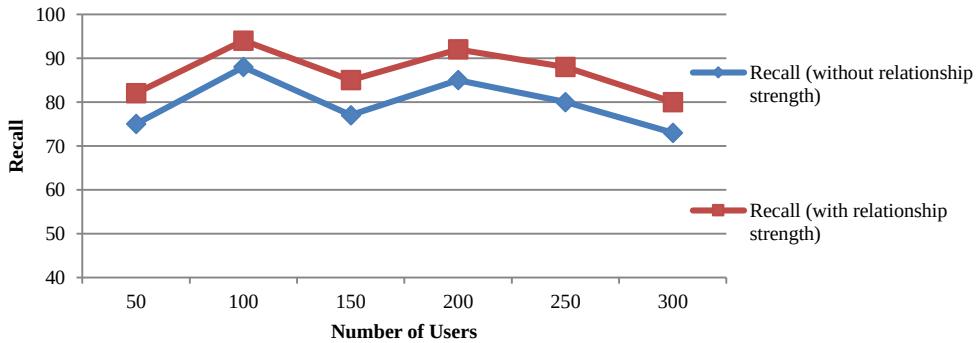

Figure 9. Comparison of Recall With Relationship Strength And Without Relationship Strength

A graph was plotted taking the total number users into consideration as the X-axis and the recall measures of correctly hiding the sensitive terms for a number of users in the Y-axis. From the graph as shown in figure 9. A comparison was done with the technique considering the relationship strength. It is shown that the technique has improved the recall to a great extent. The recall fluctuates for lengthy messages were the total number of noun phrases identified is more. But in all the cases the hiding of sensitive terms for the users were improved when the relationship strength was also taken into consideration

VI. CONCLUSION

Online social network refers to platforms where a large number of people interact among themselves by sharing information. So a huge amount of information gets shared in the OSNs each day. Hence this information is prone to privacy related issues and security breaches by various users and agents in the OSN. Hence there is a need to protect all the information being shared in the OSN. Previously various techniques have been adopted to protect the information being disclosed in the OSN. The most important of these is the trust based analysis of a user in an OSN to protect information. But with trust based analysis many limitations were observed. The first and foremost is that quantitatively evaluating the trust is not feasible in such a large network. This is mainly because the network keeps extending in size and also keeps changing rapidly. Moreover the sensitivity of the information being shared was not taken into account. Hence there was a need to analyse the sensitivity of the information being shared rather than depending wholly on the trust measure.

In order to analyse information based on its sensitivity, semantic based analysis by semantically annotating words in the information was introduced. In semantic based analysis the meaning of the information that is to be shared is analysed instead of the trust measure. Therefore this leads to a better and a more feasible technique of providing security in the OSN. To make sure that the sensitivity of the information is truly detected, the relationship strength between the users based on certain topics was identified. Among all the topics, few topics will be labelled sensitive and the user mentions the topics that are sensitive to him while registering in the OSN. This relationship strength is calculated using previously shared messages and information in order to make a justification as to whether the user has been comfortable in sharing the information with the other user in the past. Then the semantic analysis of the information that is to be shared is analysed to search for the sensitive terms which are in particular the noun phrases present in the message. Further semantic disambiguation is done to relate the exact sense of the sensitive words that were detected. Following this taxonomic categories of words were retrieved and stored in a database which has been enabled with a rule generator to facilitate the annotation process. Now the access policy is generated. The access policy and relationship strength are all compared to see if they are in accordance with the privacy settings and if it is greater than the threshold value respectively. If the conditions are all satisfied, the user has the right to view the information that was originally shared. If not, the original information is sanitized by substituting all the sensitive terms with generalized terms. Hence an automatic semantic analysis of the information shared in the social network was implemented. Experiments were conducted and the results fetched from the experiments were analysed. It was observed that the proposed system performs fairly well when compared to the existing system and also the time complexity scales linearly.

As a part of the future work, the experiments are to be extended to a large number of users and a wide variety of information and topic. Also further to facilitate the process, automatic generation of access policies are to be done with the help of the existing links and relationships in the social network.

REFERENCES

[1] Ala Atrash , Marie-Hélène Abel, Claude Moulin, "Notes and annotations as information resources in a social networking platform" Elsevier journal on Computers in Human Behavior, Vol. 51, January 2015.

[2] Abedelaziz Mohaien ,Denis Foo Kune, Eugene Y. Vasserman, Myungsun Kim," Secure Encounter-Based Mobile Social Networks: Requirements, Designs, and Tradeoffs", IEEE Transactions on Dependable and Secure Computing, Vol. 10, No. 6, October 2013.

[3] Amit Ranjbar , Muthucumaru Maheswaran, "Using community structure to control information sharing in online social networks", Elsevier journal on Computer Communication, Vol. 41 , January 2014.

[4] Anna Johnston , Salinger Privacy, Stephen Wilson," Privacy Compliance Risks for Facebook", IEEE Technology and Society Magazine, Vol. 31, No.2, June 2012.

[5] Annika Bergström" Online privacy concerns: A broad approach to understanding the concerns of different groups for different uses" Elsevier journal on Computers in Human Behavior, Vol. 53, July 2015.

[6] Bogdan Carbunar , Radu Sion , Rahul Potharaju , Moussa Ehsan," Private Badges for Geosocial Networks", IEEE Transactions on Mobile Computing, Vol. 13, No.10, August 2014.

[7] Bogdan Carbunar ; Mahmudur Rahman ; Niki Pissinou," A survey of privacy vulnerabilities and defenses in geosocial networks", IEEE Communications Magazine, Vol. 51, No. 11, November 2013.

[8] Ching-Yung Lin ,Lynn Wu , Zhen Wen , Hanghang Tong ," Social Network Analysis in Enterprise", IEEE Proceedings, Vol. 100, No. 9, August 2012.

[9] Constantinos Patsakis , Athanasios Zigomitros , Achilleas Papageorgiou , Edgar Galván-López, "Distributing privacy policies over multimedia content across multiple online social networks" Elsevier journal on Computer Networks, Vol. 75, October 2014.

[10] David Sánchez , Montserrat Batet, Alexandre Viejo, " Utility-preserving sanitization of semantically correlated terms in textual documents", Elsevier journal on Information Sciences, Vol. 279, April 2014.

[11] David Sánchez , Montserrat Batet, David Isern, Aida Valls " Ontology-based semantic similarity: A new feature-based approach " Elsevier journal on Expert Systems with Applications, Vol. 39, No. 9, July 2012

[12] David Sánchez, Montserrat Batet, and Alexandre Viejo " Automatic General-Purpose Sanitization of Textual Documents " IEEE Transactions on Information Forensics and Security, Vol. 8, No. 6, June 2013.

[13] Francesco Buccafurri,Lidia Fotia,Gianluca Lax,Vishal Saraswat, "Analysis-preserving protection of user privacy against information leakage of social-network Likes" , Elsevier journal on Information Sciences, Vol.328, September 2015.

[14] Félix Gómez Mármol, Manuel Gil Pérez, Gregorio Martínez Pérez," Reporting Offensive Content in Social Networks: Toward a Reputation-Based Assessment Approach" IEEE Tansactions on Internet Computing, Vol.18, No.2, March 2014.

[15] Hidekazu Yanagimoto, Michifumi Yoshioka, "Relationship Strength Estimation for Social Media Using Folksonomy and Network Analysis" IEEE World Congress on Computational Intelligence, June 2012.

[16] Hongxin Hu , Gail-Joon Ahn, Jan Jorgensen," Multiparty Access Control for Online Social Networks: Model and Mechanisms", IEEE Transactions on Knowledge and Data Engineering ,Vol.25 , No.7, May 2013.

[17] Jin Chen, Jerry Wenjie Ping, Yunjie Calvin Xu, Bernard C. Y. Tan "Information Privacy Concern About Peer Disclosure in Online Social Networks" IEEE Transactions on Engineering Management , Vol. 62, No. 3, August 2015.

[18] Jun Zhou, Zhenfu Cao, Xiaolei Dong, Xiaodong Lin," Securing m-healthcare social networks: challenges, countermeasures and future directions", Vol. 20, No. 4, August 2013.

[19] Kai Li, Zhangxi Lin, Xiaowen Wang, "An empirical analysis of users' privacy disclosure behaviours on social network sites ", Elsevier journal on Information and Management, Vol. 52, No. 7, November 2015.

[20] Kaze Wong, Angus Wong, Su-Kit Tang, Alan Yeung and Wei Fan ,"Trust and Privacy Exploitation in Online Social Networks" IEEE Transaction on Computer Society, Vol. 16, No. 5, October 2014.

[21] Lo-Yao Yeh, Yu-Lun Huang, Anthony D. Joseph, Shiuhpyng Winston Shieh," A Batch-Authenticated and Key Agreement Framework for P2P-Based Online Social Networks",IEEE Tansactions on Vehicula, Vol. 61, No. 4, May 2012.

[22] Lorenz Schwittmann, Matthaus Wander, Christopher Boelmann, Torben Weis, " Privacy Preservation in Decentralized Online Social Networks" IEEE transactions on Internet Computing, December 2013.

[23] Lorrie Faith Cranor, Norman Sadeh," Privacy Engineering Emerges as a Hot New Career", IEEE Potentials, Vol. 32, No.6, November 2013.

[24] Malik Imran-Daud,David Sanchez,Alexandre Viejo, " Privacy driven access control in social networks by means of automatic semantic annotation", Elsevier journal on Computer Communications, January 2016.

[25] Michael Fire , Beer-Sheva, Israel,Roy Goldschmidt,Yuval Elovici" Online Social Networks: Threats and Solutions", IEEE Journal on Communications Survey and Tutorial, Vol. 16, No.4, November 2014.

[26] Miriam Bartsch , Tobias Dienlin ," Control your Facebook: An analysis of online privacy literacy" Elsevier journal on Computers in Human Behavior, Vol. 56, December 2015.

[27] Nan Shen , Jun Yang , Ke Yuanb, Chuan Fua, Chunfu Jia " An efficient and privacy-preserving location sharing mechanism" Elsevier journal on Computer Standards & Interfaces, Vol. 44, June 2015.

[28] Natalia Criadoa, Jose M. Such, "Implicit Contextual Integrity in Online Social Networks" Elsevier journal on Information Sciences, Vol. 325, July 2015.

[29] Raymond Heatherly, Murat Kantarcioglu, Bhavani Thuraisingham," Preventing Private Information Inference Attacks on Social Networks",IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 8, June 2013.

[30] Rudi L. Cilibrasi and Paul M.B. Vitanyi "The Google Similarity Distance", IEEE Transactions on Knowledge and Data Engineering , Vol. 19, No. 3, March 2007.

[31] Seda Gürses, Claudia Diaz," Two tales of privacy in online social networks", IEEE Journal on Security and Privacy , Vol.11, No. 3, May 2013.

[32] Sami Vihavainen , Airi Lampinen , Antti Oulasvirta , Suvi Silfverberg," The Clash between Privacy and Automation in Social Media", IEEE Tansactions on Pervasive Computing, Vol.13, No.1, January 2013.

[33] Tabitha L. James , Merrill Warkentin , Stephane E. Collignon, "A dual privacy decision model for online social networks", Elsevier journal on Information and Management, Vol. 52, August 2015.

[34] Xi Chen, Katina Michael," Privacy Issues and Solutions in Social Network Sites", IEEE Technology and Society Magazine, Vol. 31, No.4, December 2012.

[35] Yongbo Zeng, Liudong Xing, "A Study of Online Social Network Privacy Via the TAPE Framework", IEEE Journal in Signal Processing, Vol. 9, No. 7, October 2015.